WESTERN POWER DISTRIBUTION

INNOVATION

NEXT GENERATION NETWORKS

LCT DETECTION

CLOSEDOWN REPORT

Report Title	:	LCT DETECTION CLOSEDOWN REPORT
Report Status	:	DRAFT
Project Reference:		WPD_NIA_036
Date	:	28.02.2019

Document Control

	Name	Date
Prepared by:	Gill Nowell, ElectraLink	25.02.2019
Reviewed by:	Dan Hopkinson, ElectraLink	28.02.2019
Approved by:	Jon Berry	16.07.2019

Revision History

Date	Issue	Status
28.02.2019	1.0	Draft
08.04.2019	1.1	Draft
18.04.2019	1.2	Issue for Approval
24.05.2019	1.3	Revised issue for Approval
06.06.2019	2.0	Issue for Approval

Contents

Executive Summary

The Low Carbon Technologies (LCT) Detection project has been managed and delivered by ElectraLink, the UK's Energy Market Data Hub, in partnership with IBM. It has been funded by Western Power Distribution's (WPD) Network Innovation Allowance. The driver for the project was the need for WPD to gain better insight into where electric vehicles (EVs) and LCTs such as solar panels and heat pumps are connected to its Low Voltage network, at a domestic level. Better visibility is essential in order for WPD, and other Distribution Network Operators, to manage their networks effectively as instances of such new demand and generation start to proliferate.

The project has combined over six years' worth of structured and freeform data from ElectraLink's Data Transfer Service dataset, with WPD's asset database and used machine learning and cognitive analysis to identify previously unknown instances of EVs and LCTs. An agile sprint process was used to manipulate the data, train and build the proof of concept model. A 'design thinking workshop' at project outset engaged relevant personnel across WPD's business areas, to ensure that the project output was designed to meet WPD's business needs. An additional business values report was delivered mid-way through the sprint process to ensure continuing alignment with WPD's business requirements.

The main project output is two proof of concept models that have successfully identified unregistered EVs and LCTs on the network – from both structured and unstructured data.

The models have found indications of 15,000 previously unknown EVs and solar panels connected to WPD's local electricity network. The data suggests that there could be 13% more households with electric vehicles and solar panels on WPD's network than was previously thought.

The project has also revealed valuable insights around energy consumption in general, including a 25% reduction in domestic electricity usage following solar panel installation, as well as a 5% increase in energy consumption where electric vehicle charge points are installed.

The proportion of LCTs connected to the low voltage network is high in rural areas considering the density of the population. The findings also show that EVs and solar panels are more prevalent in affluent areas while solar panels are also present in areas of high deprivation, likely due to leasing of social housing roof space for solar panels.

The proof of concept model developed under this project has been based on consumption data only. It is clear from the static analyses carried out on the aggregated data that bringing in socio-economic data will enhance the model. Another key piece of learning centres on the need for more granular data and a negative dataset; the project developed its own negative data set which precluded use of, for example, socio-economic data. The proof of concept model under the LCT Detection project is an essential stepping stone to development of an holistic virtual monitoring capability for WPD to underpin its transition to Distribution System Operator.

The total project cost was £346,020, of which £311,418 was NIA. The project was delivered to time and on budget.

1 Project Background

ElectraLink, the UK's Energy Market Data Hub (EMDH), has partnered with Western Power Distribution (WPD), the company responsible for electricity distribution in South Wales, the South West and Midlands, on a ground-breaking project that enables WPD to identify unregistered electric vehicles and other low carbon technologies on its network. The project has been funded through WPD's Network Innovation Allowance.

The energy market is complex and evolving, particularly with growing smart technologies and embedded, renewable generation. For DNOs, the increasing number of 'invisible' changes (growth of Electric Vehicles and Embedded Generation) challenge existing network practices to the extent that the status quo is no longer possible.

Managing this new demand will require smart charging and other smart solutions but these necessitate visibility of where Low Carbon Technologies (LCT) are connected to the distribution network at the local level – something that, up to now, has proved difficult for Distribution Network Operators (DNOs, the companies responsible for managing the networks that deliver electricity at a local level).

The LCT Detection project set out to help WPD identify areas where there is a high proliferation of electric vehicles, solar panels or other low carbon technologies and so manage its network accordingly.

The project has made use of ElectraLink's unique energy market dataset combined with other data and enhanced by IBM's Cognitive computing capability to provide WPD with a proof of concept model that has the potential to provide much needed visibility of the ever-growing demand for electric vehicles, as well as other LCTs such as solar panels and heat pumps. By tying together data that had not been combined previously – including text – this has provided new insights into where unregistered LCT equipment is located.

Going forward, this will allow WPD's network planners to better assess the existing network capacity and understand how this is likely to change in the future. This will have clear benefits for customers through potential for avoidance or reduction of reinforcement costs, and less associated disruption.

2 Scope and Objectives

By using innovative data analytic techniques, this project tackles a key network and operational issue which forms a part of an overarching industry need – the increased requirement for data to support energy market operations. This project has taken industry data from the Data Transfer Service (DTS) data set and applied leading-edge cognitive analytics to provide WPD with a proof of concept tool. This tool has the potential for improved visibility of EVs and LCTs to support forecasting of the proliferation of EVs and solar panels across networks to support network planning, including the options of active/flexible network management. The project has helped define the requirements for the delivery of an enhanced dataset proof of concept model allowing us to leverage the analytics tools and techniques to support WPD to identify unregistered EVs and LCT, and to understand how best to validate suspected installations. The project has analysed data on a number of representative network topologies across WPD's Electricity Service Areas (ESAs). All ESAs were required in order to ensure a sufficient number of known locations to help train and validate the model. This is particularly true for heat pumps where there are relatively few records, as was found to be the case. The number of customers assessed for LCT identification reflected the volumes required to generate a sufficiently large candidate set to validate the model and to identify regional differences in model effectiveness. Many of the project costs were not scale dependent.

By using ElectraLink's DTS dataset, combining this with a range of other structured and unstructured data and then applying IBM's Cognitive analytics, the objective was to identify patterns in the data that indicate the presence of EV, solar panels or other LCTs that had not previously been identified. IBM applied its Watson technology to perform advanced analytics on the ElectraLink data, combined with other datasets. IBM used a progressive and iterative methodology to detect patterns in the data that was not detected hitherto. By improving detection of EVs and LCTs on the network, the project has built the foundation for improving forecasting capabilities and, ultimately, garner an understanding the effectiveness and costs for the various options would allow for the validation process to be optimised.

Objective	Status
By using ElectraLink's DTS dataset, combining this with a range of other structured and unstructured data and then applying IBM's Cognitive analytics, the objective was to identify patterns in the data that indicate the presence of EV, solar panels or other LCTs that had not previously been identified.	✓

Table 1-1 Objectives

3 Success Criteria

Success Criteria	Status
An advanced PoC analytics model that identifies LCT on the LV network	✓
Introduction of the project to a DNO and energy industry audience at WPD's Balancing Act event in November 2018	✓
Presentation of the final report to a DNO and industry audience at WPD's Balancing Act event in May / June 2019	The project was disseminated via a webinar to an audience of c. 60 industry delegates.
Analysis of data on a number of representative network topologies across WPD's Electricity Service Areas (ESAs)	✓
Validation – recommendations as to what validation approach WPD should use for each candidate set	✓
Delivery to WPD of a PoC model - a process design document and demonstration dashboard that will identify Low Carbon Technologies (LCTs) to support network planning and investment strategy	✓

Table 1-2 Success Criteria

4 Details of Work Carried Out

4.1 Project Plan

Due to the iterative nature of the CRISP-DM Cycle (described below), it was decided to execute the project using the concept of "Sprints" borrowed from the Agile project management approach. This approach enabled us to deliver the model and engage with key stakeholders at frequent intervals as described in the Milestone Plan:

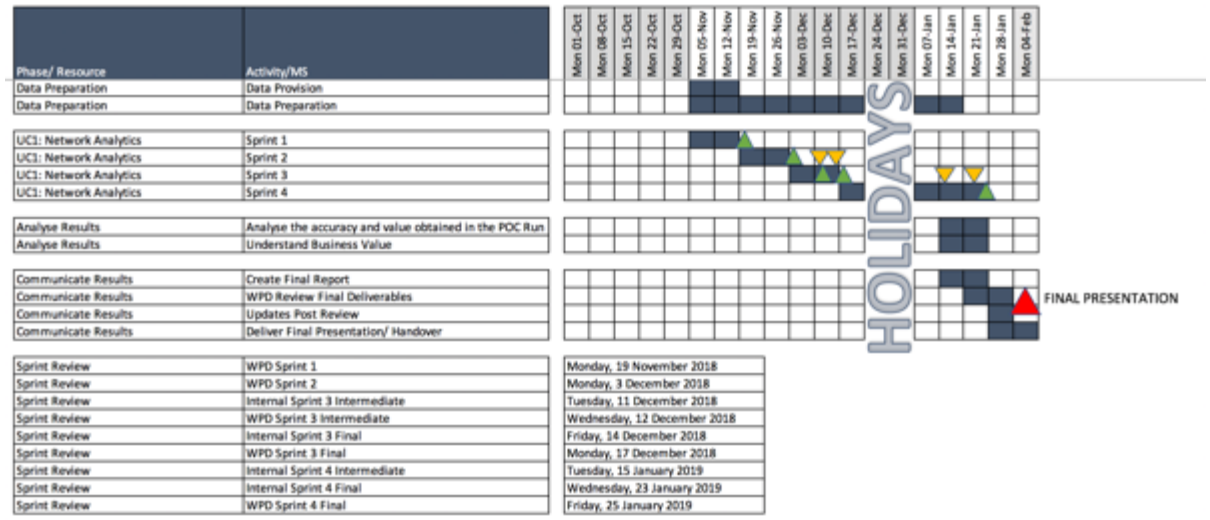


Figure 4-1 Sprint Plan & Key Milestones

Four sprints were planned with high level tasks broken out and allocated to each sprint.

Sprint 1	Complete Workshop Outputs
Sprint 1	Summarise Data Sets
Sprint 1	Complete Environment Setup
Sprint 1	Data Upload
Sprint 1	Data Summarised
Sprint 1	Begin Initial Data Cleansing/ Wrangling
Sprint 2	Document Business Value
Sprint 2	Complete Data Quality Assessment
Sprint 2	Define Modelling Approach
Sprint 2	Complete Initial Data Cleansing/ Wrangling
Sprint	Data Analysis

3	
Sprint 3	Data Insights
Sprint 3	Initial Models for Analysis
Sprint 3	Initial Feature Engineering
Sprint 4	Final Model
Sprint 4	Learnings
Sprint 4	Draft Final Report
Sprint 4	Final Dashboard Presentation (static mock-up)

Table 4-1 Sprint Tasks

Detailed tasks were tracked on a Kanban board using Trello software:

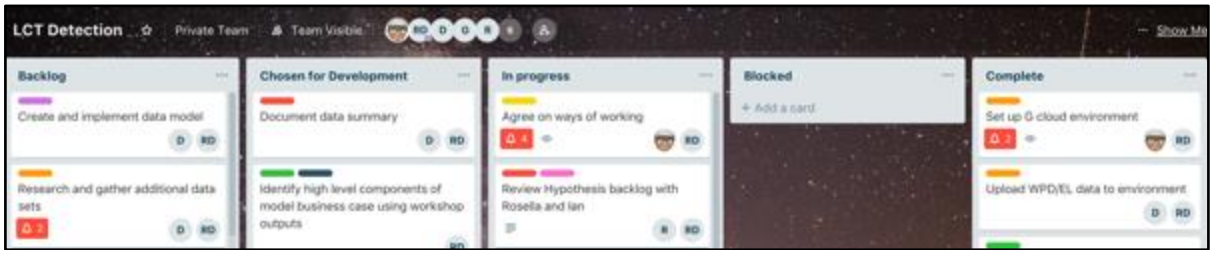


Figure 4-2 Screenshot of Project Trello Board

The Kanban board gave a clear view of tasks that were:

- Backlog: the items yet to be allocated (future sprints);
- Chosen for Development: in current sprint but not yet started;
- In Progress: currently being worked on;
- Blocked: a problem that requires escalation through PM;
- Complete.

Tasks were picked up and completed in a flexible manner that allowed the team to complete a task and then move onto the next.

The team met daily for a brief (15 minutes) “Stand-Up” call based on the Kanban board – where progress and next steps are briefly discussed and “Blockers” to progress called out to the Project Managers for immediate and targeted escalation/ resolution. This model of project management supported the success of the project and will be highlighted in the Lessons Learnt in this document.

4.2 Overview of Methodology

The LCT Detection project is a data science project which is executed based on the widely accepted CRISP-DM cycle:

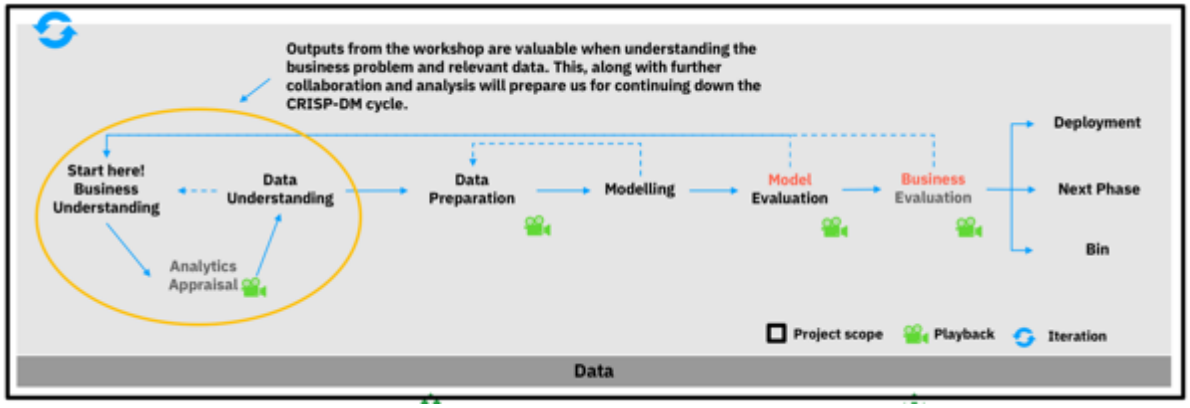


Figure 4-3 CRISP-DM Cycle

4.2.1 Business Understanding / Data Understanding– Design Thinking

The data analytics phase of the project was initiated on Friday 2nd November 2018 with a Design Thinking Workshop led by IBM and attended by WPD and ElectraLink. The purpose of the workshop was to bring together WPD, ElectraLink and IBM to provide a shared understanding of end-users and their current processes related to LCT proliferation, to enable the development of outputs designed with users in mind. Additionally, the workshop captured business-led hypotheses around LCT proliferation and identified data that can be used to investigate them.

Throughout the day participants were guided through different tasks and asked to focus on the challenge for the day:

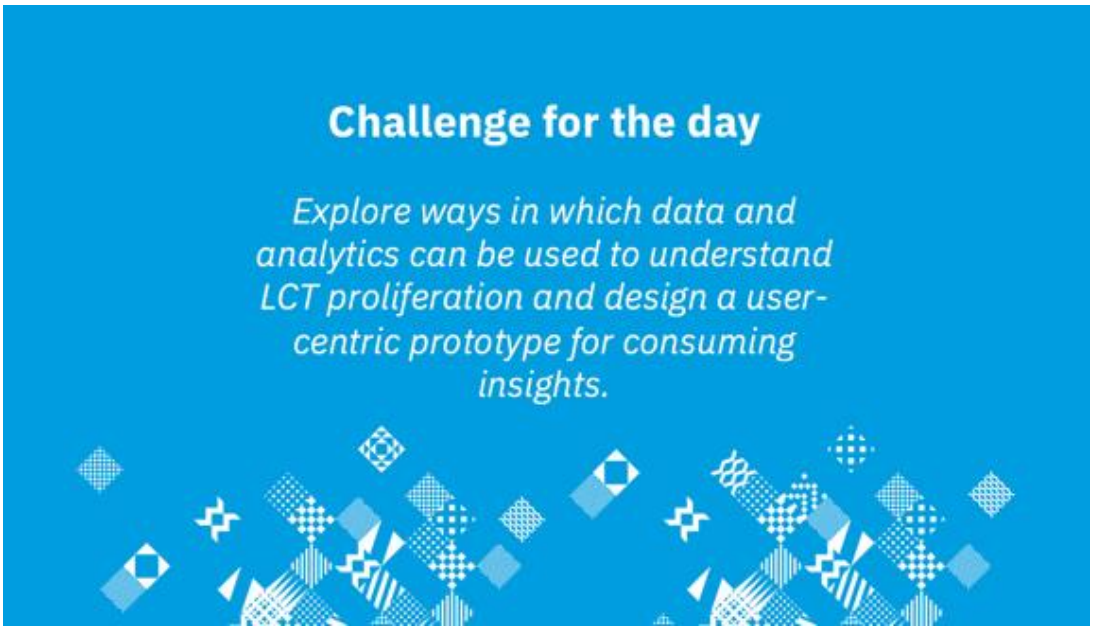


Figure 4-4 Challenge for the Day

4.2.2 Output Summary

Common themes appeared in the hypotheses around LCT proliferation across the groups. In general hypotheses were in one of the following formats:

- Consumption at a Meter Point Administration point (MPAN) level will change following the installation of LCT at that property;
- It is likely LCT can be identified at an MPAN level by classifying unstructured notes related to that property;
- The demographics of an area will influence LCT proliferation at an MPAN level;
- Features/characteristics of a home will influence LCT proliferation at an MPAN level.

In total 25 distinct hypotheses were generated. For each hypothesis an approximate number for expected value and feasibility out of 10 (10 being most valuable/feasible) has been recorded. The specific data points required, and data sets identified in the workshop to investigate the hypothesis have been listed. The data status has been labelled as one of the following:

- Obtained: If the data is included in the WPD/ElectraLink data sets already provided;
- Sources Identified: If available sources of this data are known but need to be gathered and assessed for suitability of inclusion;
- Research required: If research is required to understand if such data exists and it is reasonable to gather and format as part of the Proof of Concept (PoC).

Below is a table of the top hypotheses. In general hypotheses from site visits notes, relating to the demographics of an area or related to changes in consumption were perceived as more value and feasible compared to hypotheses related to features of a property.

Hypothesis	Feasibility	Value	Type of LCT	Data Points Required	Data Sources	Data Status	Average score
The presence of LCT at an MPAN level could be included in site visit notes	7	9	All	Notes classified as mentioning LCT or not	Site visit data	Obtained	8
Homes in areas with a high deprivation index are less likely to have LCT	8	8	All	Deprivation index at a postcode level	Publicly available deprivation index	Sources identified	8
Homes in rural areas are less likely to have LCT	8	8	All	A flag for whether a home is rural, suburban or urban	Use sources of open data	Sources identified	8
If meter readings a show significant drop over a long period of time and the tenant did not change, then it is likely PV has been installed *Note: weather could be a factor in consumption drop observed*	8	7	PV	Meter readings, Change of tenancy indicators at MPAN level, Time of LCT installation for known instances, *Measure of sunlight*	Meter readings, Change of tenancy indicators, Known LCT *Weather data*	Obtained: Meter reading, change of tenancy, Known LCT Sources identified: weather	7.5
There will be a step change in consumption after and LCT is installed	8	7	All	Meter readings, Change of tenancy indicators at MPAN level, Time of LCT installation for known instances	Meter readings, Change of tenancy indicators, Known LCT	Obtained	7.5
EV is more likely in affluent areas	7	8	EV	Indicators of affluence (such as council tax band)	Affluence data that can be linked to postcode	Sources identified	7.5
LCT unlikely in properties with pre-payment meters	8	6	All	Meter type	Meter details	Obtained	7
Homes in areas with public EV charging points are more likely to have EV and hence personal EV chargers	6	7	EV	Location of public charging points	Charging point record	Research required	6.5
LCT more likely in for certain council tax bands	6	7	All	Council house band at an MPAN level	Council records	Research required	6.5

Table 4-2 Hypotheses

4.3 Data Preparation/ Modelling/ Evaluation

The bulk of the project effort was centred around the preparation of data and subsequent modelling and evaluation. The full details of each sprint are presented in the Sprint Output Reports that have been delivered to WPD, and which are summarised below.

4.4 Sprint 1 Summary

Sprint 1 included tasks to establish environments, summarise the Design Thinking workshop, summarise the data and data sets available and complete data upload. All tasks were completed within the allocated time.

4.4.1 Environment Set Up

The environment selected to carry out the data analytics work is IBM Watson Data Studio. This environment is cloud based and offers a catalogue of components that can be selected and combined to support the tasks required. For this project we selected:

- IBM Cloud Object Storage – allows large datasets to be uploaded, stored, accessed and written.
- Jupyter Notebooks (used for writing and testing code) have been created using Python 3.5
- PySpark - a tool for analysing large data sets efficiently - chosen due to the size of the Meter Readings table (7GB+ size, 100 million+ records).

4.4.2 Design Thinking Workshop Outputs

- The outputs of the Design Thinking Workshop were:
  - user profiles
  - current processes & associated pain points / opportunities
  - user needs statements
  - hypotheses (along with feasibility, value, type of LCT concerned, and additional comments from attendees)
  - possible additional external datasets
  - prototype designs

4.4.3 Summary of Data Sets

The following data sets were uploaded ready for analysis:

Known LCT		
Field	**Description**	**Format**
pMPAN	Psuedonymised version of Meter Point Adminstration Number	Varchar
LCT_TYPE	Low Carbon Technology installed at Metering Point	Varchar
CAPACITY	the demand/export capacity of the technology (kW)	Float
CommissionedDate	Date at which LCT was connected	Date

Known Generation		
Field	**Description**	**Format**
pMPAN	Psuedonymised version of Meter Point Adminstration Number	Varchar
GEN_TYPE	Type of embedded generation installed at Metering Point	Varchar
CAPACITY	the export capacity of the generator (kW)	Float
Commissioning_Date	Date at which generator was connected	Date

Table 4-3 - Known instance of LCT

Due to data privacy, IBM was given the pMPAN for every household and location, which refers to the pseudonymised MPAN. The datasets were then uploaded.

4.4.4 Data Summarised

The DTS and CROWN datasets listed above were uploaded and joined to understand the data that IBM would be analysing. This data was summarised as part of data understanding, to

check data quality and to understand how data could be joined for modelling. All data sets relate to MPANs in WPD's East Midlands region.

Key findings included:

- Analysis of the known instances of LCT;
- Analysis of MPANs contained in each data set;
- Time distribution of meter readings;
- Analysis of reading types;
- Meter installation date;
- Distribution of meter types (example below).

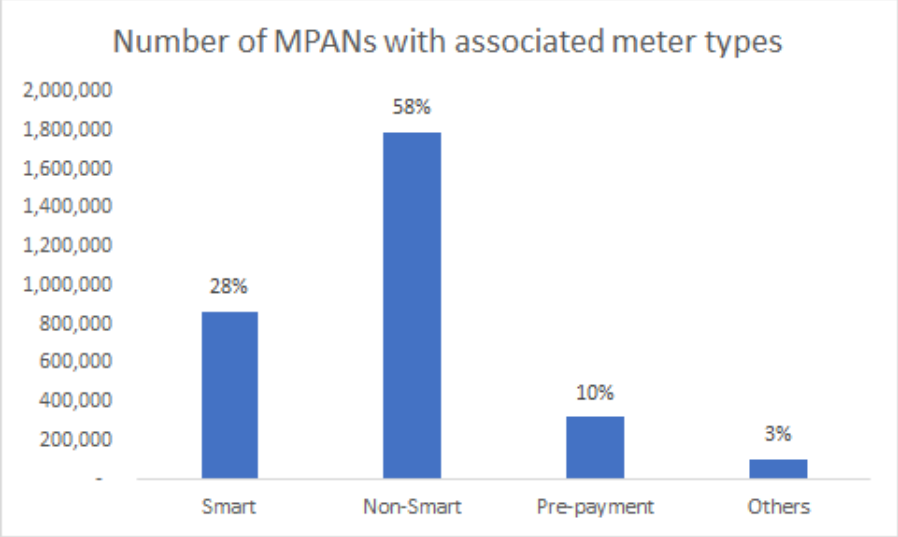


Figure 4-5 - Number of MPANs by Meter Type

4.5 Sprint 2 Summary

Sprint 2 included tasks to Document Business Value, Complete Data Quality Assessment, Define Modelling Approach, Complete Initial Data Cleansing/ Wrangling.

4.5.1 Business Value

The Business Value approach was clarified during Sprint 2. Business Value would be delivered through creation of a PoC that could identify potential LCT and/ or EV presence on the LV network. This objective could be fulfilled by providing the following data and features per MPAN record to help define the PoC:

1. MPAN
2. Substation
3. LCT? (Yes/No)
4. If yes, LCT type
5. Time installed
6. Regional information (Urban/Suburban/Rural or Deprivation Index)
7. Last EAC read
8. Identification type (how the project has identified the LCT – i.e. through consumption values or free text analysis)

4.5.2 Data Quality Assessment

Training Data:

In order to find LCTs, in Sprint 2 we analysed whether we had sufficient information about LCTs to find them. Whilst we had complete access to the DTS dataset, the better the information/data that is obtained about an LCT installation, the more uncertainty is reduced. There two key types of datasets that are integral to this project and these are the positive and negative datasets.

Having a negative dataset helps reduce the risk of normal changes in consumption, such as seasonal changes, being used as evidence of an LCT.

There was an important decision/ turning point in the project which arose during Sprint 2. In order to train a model, both positive and negative known instances of LCT must be present. E.g., Properties a, b, c and d have PV installed; properties e, f, g and h DO NOT have PV installed. The reason you need both datasets is that the negative dataset will help the model understand what 'normal' non-LCT households look like, as a comparator to understand what stands out for LCT households. Whilst we had a positive dataset from the CROWN dataset, we did not have a negative dataset available; therefore, we had to use the data available to find a negative dataset to ensure the success of the model. To tackle this issue of no data for negative instances, the project was faced with two options:

- Create a negative set from the known instances of LCT by taking periods from before LCT was installed as a negative set. Features can be created that measure normalised change in consumption compared for each MPAN. The model can then learn to classify periods of consumption for an MPAN based on how consumption has changed;
- Assume certain MPANs do not have LCTs and use this as a negative set. These assumptions could be made by using the hypotheses from the workshop, for example properties in areas of high deprivation or with pre-payment meters do not have LCT. Additionally, analysis would need to be done to check that properties with these features do not have markedly different consumption patterns than the general population.

The selected option was to create a negative set from the known instances of LCT by taking periods from before LCT was installed as a negative set. Features can be created that measure normalized change in consumption compared for each MPAN. The model can then learn to classify periods of consumption for an MPAN based on how consumption has changed. The reason for this, as will be explained in the Sprint 3 analysis, is that there is no definitive dataset to prove that low deprivation areas do not have LCTs (in fact the opposite is true for PV), whilst other hypothesis, such as the presence of a prepayment meter will indicate no LCT, is also not correct.

Features about the properties the MPAN belongs to (such as demographics of the area or type of home) cannot be used in this model, as there is no information about properties without LCT for the model to learn from.

The training data set was expanded from the East Midlands and taken from all WPD regions and cleansed to remove erroneous or dubious data items. When these records are removed, the following training set sizes remain:

LCT Type	# of MPANs	% of MPANs
Photovoltaic	123,048	89.4%
EV Charge Point	11,470	8.3%
Other (including heat pump)	3,150	2.3%

Table 4-4 - Training set sizes by LCT type after dropping invalid records

Unstructured Data

In Sprint 2 there was a key decision point around the treatment of unstructured data. The unstructured dataset contains ~7.5million distinct rows for ~2.5million MPANs.

Typically, the benefit gained from training a natural language classifier is that it can identify information that has been expressed in nuanced ways. Short comments mean it is difficult for a trained classifier to be more informative than a comprehensive key word search. Therefore, it was decided to create a comprehensive key word search in order to identify LCT presence.

13% of the comments are one of the following:

- 'No Access'
- 'K'
- blank

A further 11% are:

- 'NRQ'
- 'No answer'
- 'white door'
- 'Unable to obtain meter reading'
- 'No more info<>'.

None of these indicate any LCT. All other comments are less common.

After removing these values, 43% of the remaining comments have 3 or fewer words. A 3-word comment means there are only a limited number of ways that LCT could be mentioned and there is less opportunity for the nuances that are often present in natural language.

Other Data Sets

Other data sets considered were:

- Meter Details;
- Change of Tenancy;
- Profile Class; and
- Demographic data.

These datasets were joined for all the MPANs for analysis. As the key identifier for the consumption modelling is a change in consumption, any MPAN that had a Change of Tennant (CoT) flag at the same time as a consumption change was not included in the modelling.

Details of the analysis are contained in the Sprint 2 Output Status Report, that has been delivered to WPD.

4.5.3 Modelling Approach

Overview

Hypotheses around LCT proliferation had the following key themes:

- Consumption: LCT installation leads to a change in consumption;
- Unstructured data: Mentions of LCT can found and classified from the unstructured site visit data;
- Demographics: LCT is more likely in MPANs within a certain property type or in an area with certain demographics.

The modelling approach is to build a “consumption” model that identifies whether an LCT is present at an MPAN level by learning from the differences in consumption patterns when LCT is present compared to when it is not present. Additionally, a model will be built that classifies unstructured text to identify whether or not it mentions LCT presence at an MPAN level.

Each MPAN will be given a classification by both the consumption model and the unstructured model. This will be combined in the final output so that for each MPAN it shows whether the MPAN has been classified as having LCT present and which model has given it this classification.

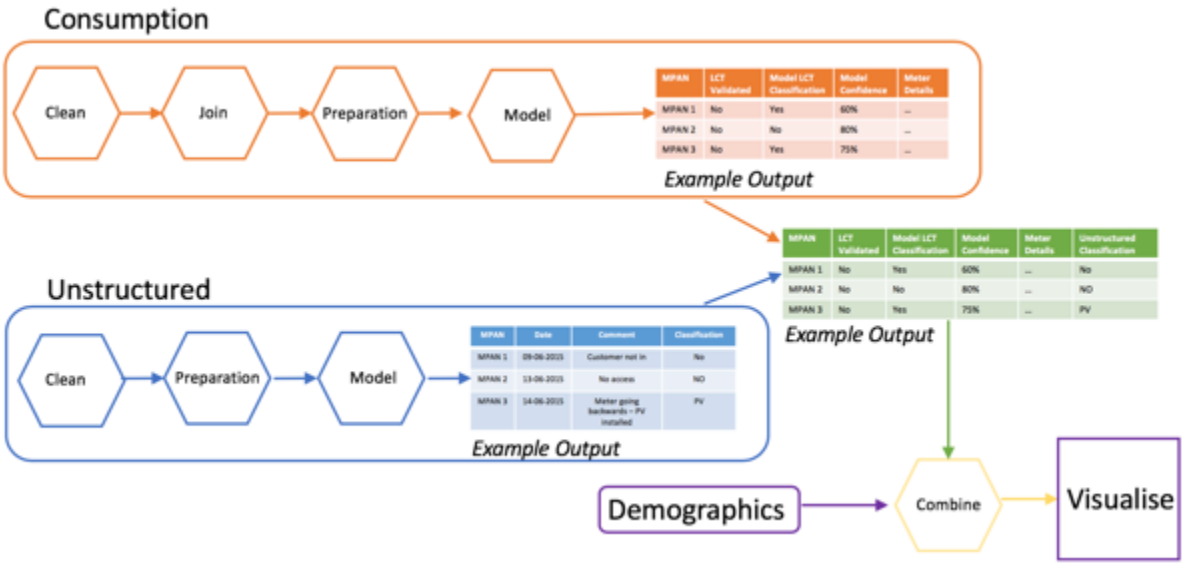


Figure 4-6 - Model Architecture

Due to the one-sided nature of the training data, demographic hypotheses cannot be included in the model but can be added to the output and visualised which could provide insight for inferences that need to be investigated.

4.5.4 Initial Data Cleansing/ Wrangling

Overview

In Sprint 2 an architecture of the modules of the consumption model was developed:

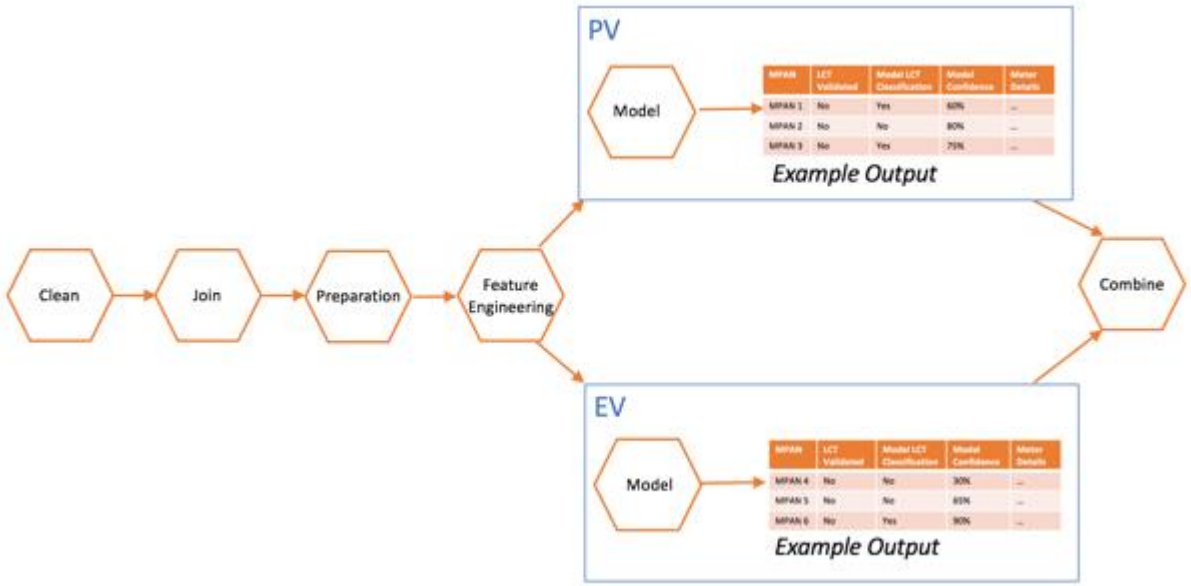


Figure 4-7 - LCT Classification Model Architecture

The data was cleaned, joined and prepared before being sent to models that classify LCT presence at an MPAN level.

4.6 Sprint 3 Summary

There were some overrunning tasks in Sprint 2, so it was decided to update the Sprint Plan as follows with Sprint 3 running across the holiday period completing on 11th January.

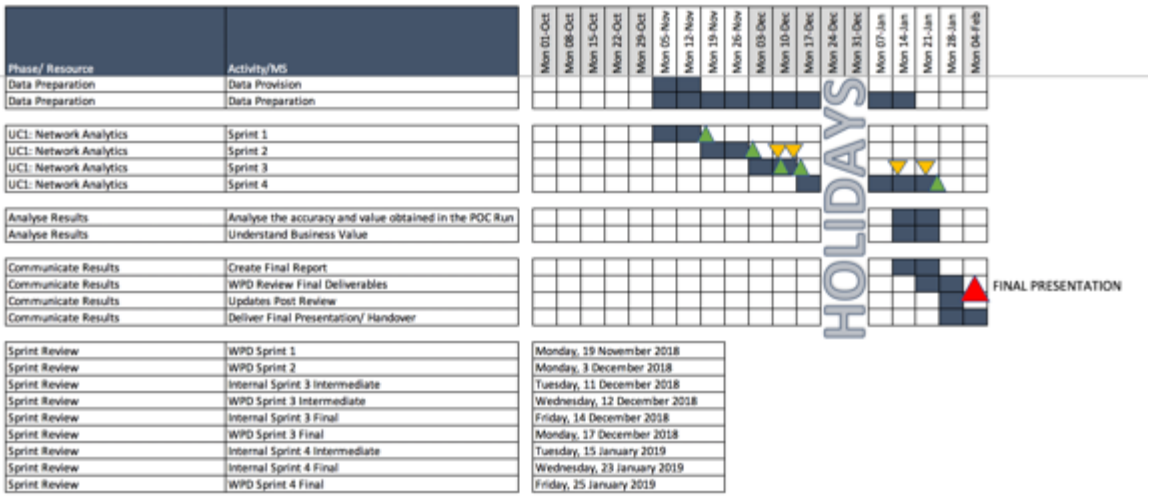


Figure 4-8 - Updated Sprint Plan & Key Milestones

The bulk of the work carried out in sprint 3 was around analysis of the data we have from the DTS and other data sets.

4.6.1 Data Analysis & Insights

There were some key findings in Sprint 3 which reinforced the hypothesis that we would find indicators of PV/ EV presence using consumption data. Details of the analysis are contained in the Sprint 3 Output Status Report, as delivered to WPD under the project.

Demographic Analysis: Introduction

Hypotheses related to LCT proliferation have been investigated. This is to give a deeper understanding of LCT proliferation prior to modelling and inform feature engineering. Specifically, the following analysis was performed, comparing the differences for MPANs where LCT is present to the entire population:

- The proportion of MPANs with each meter type
- The proportion of MPANs in rural and urban areas
- The distribution of MPANs in areas of low or high deprivation

The analysis has been repeated for domestic MPANs, non-domestic MPANs and MPANs where the profile class is unknown.

~78% of MPANs in the dataset are domestic, ~8% are non-domestic and ~14% are unknown.

Known instances of the following LCT types have been analysed:

- PV (~123k MPANs, 1.8% of total MPANs)
- EV (~11k MPANs, 0.2% of total MPANs)
- Heat Pumps (~1.2k MPANs, 0.02% of total MPANs)

Other LCTs have been excluded from analysis due to low volumes of known instances.

Demographic Analysis: Summary of Results

As hypothesised, EV proliferation is skewed towards more affluent areas. PV also shows peaks for more affluent households. We also see peaks for PV at the less affluent end of the scale which may be linked to our hypothesis that less affluent areas with social housing is more likely to have PV or be participating in a rent-a-roof scheme.

Results: Meter Type

The data shows us the progressive uptake of LCT with the average installation dates being: PV – 2013, EV Chargers – 2016, Heat Pumps – 2017.
For each profile class the proportion of all MPANs of each meter types have been analysed. Initially this was compared to the proportion of MPANs with LCT present at the time of LCT installation.

The average installation date for known PV is 08/09/2013, for known EV is 26/06/2016 and for Heat pumps (HP) is 24/09/2017.

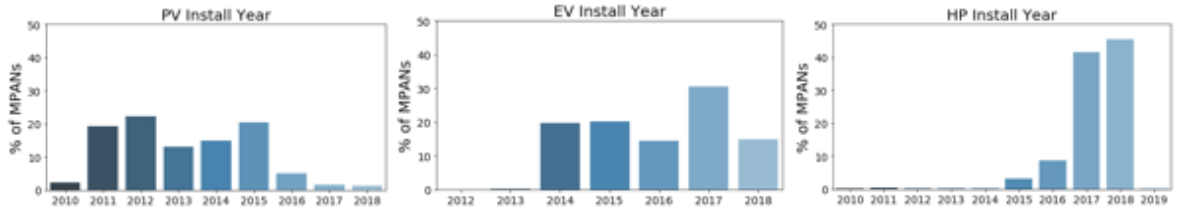


Figure 4-9 - Installation Years

Looking at all domestic MPANs – 9% have a pre-payment meter. In comparison, just 1% of domestic MPANs with known EV have a pre-payment meter: indicating a low uptake of EV in pre-payment meter properties. For PV, over 7% of domestic properties with known PV have a pre-payment meter: indicating an even uptake of PV in pre-payment meter properties. Heat pumps are more prevalent in pre-payment meter properties with over 17% of known heat pumps installed with a pre-payment meter.

Non-domestic MPANs are much more likely to have an "other" meter type when compared with domestic MPANs. Meter types have been classified as "other" if they are not pre-payment, smart or regular non-smart meters.

Results: Rural/Urban Analysis

ElectraLink provided data categorising the location of MPANs. The urban/rural split has been analysed for all MPANs in the data and for MPANs with LCT present.

Analysis shows that there is an increase in LCTs present in settlements falling into the "village" category (E1). Therefore, the hypothesis that rural areas are less likely to have LCTs is not correct.

Classification of Rural and Urban Settlements

Settlement data has been classified according to the ONS categories as detailed below:

Urban/ Rural	Sparse/ Not Sparse	Settlement	Category Code
Urban	Not Sparse Setting	Major Conurbation	A1
Urban	Not Sparse Setting	Minor Conurbation	B1
Urban	Not Sparse Setting	City & Town	C1
Urban	Sparse Setting	City & Town	C2
Rural	Not Sparse Setting	Town & Fringe	D1
Rural	Not Sparse Setting	Villages	E1
Rural	Not Sparse Setting	Hamlets & Isolated Dwellings	F1
Rural	Sparse Setting	Town & Fringe	D2
Rural	Sparse Setting	Villages	E2
Rural	Sparse Setting	Hamlets & Isolated Dwellings	F2

Table 4-5 Rural Urban Classification

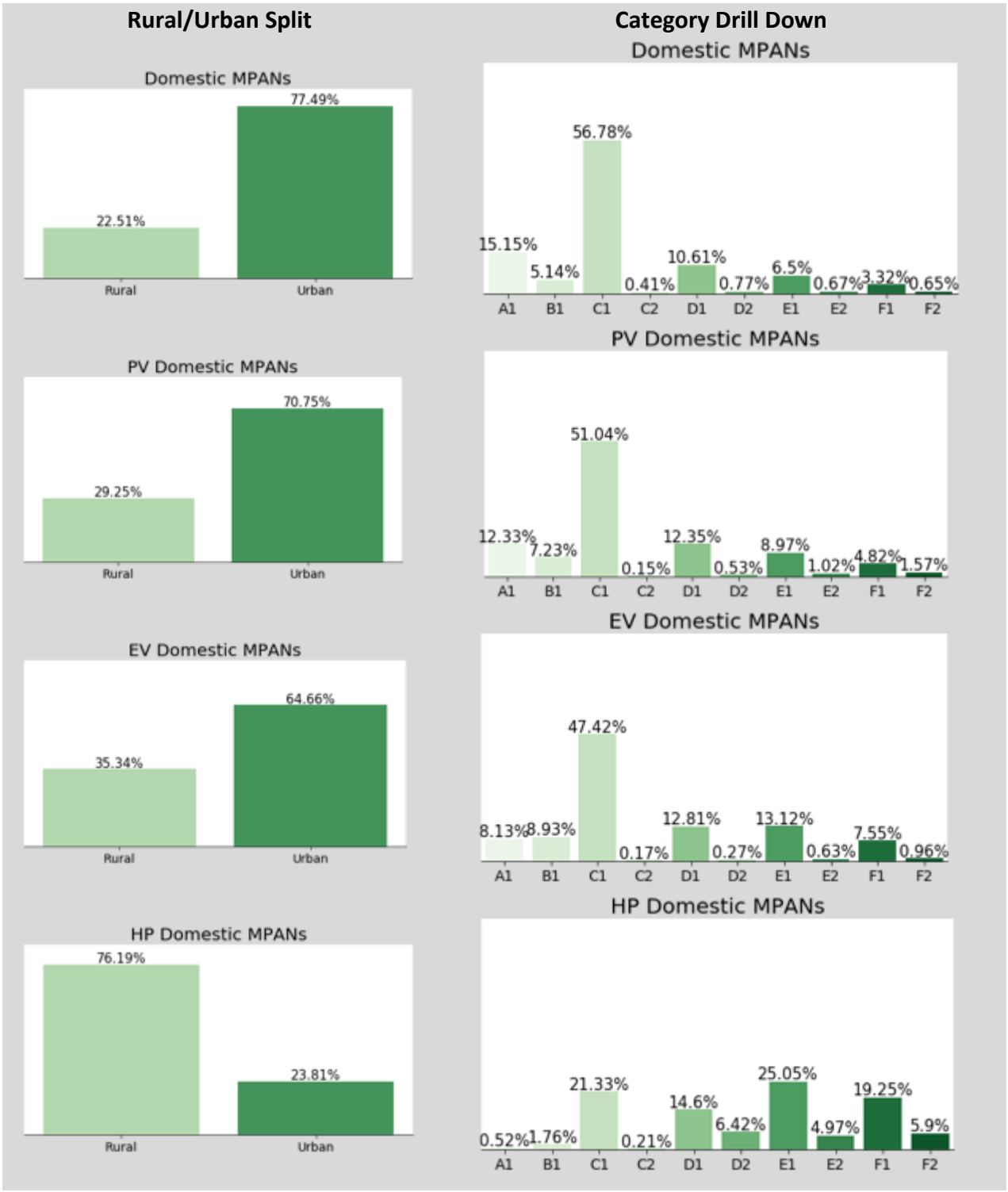


Figure 4-10 - Rural Urban Split - Domestic

For non-domestic MPANs, there is a larger proportional increase of rural MPANs when LCT present compared to domestic MPANs. A large increase is seen in the Hamlet/dwelling category (F1).

For MPANs of unknown class, again there is an increase in the proportion of MPANs in rural areas when LCT is present.

Results: Deprivation Index

The Deprivation Index categorises areas based on their deprivation, with 1 being most deprived. England and Wales are indexed differently, so analysis has been done on each area separately – comparing deprivation for the entire population and comparing this with properties with LCT.

England Analysis

Average deprivation was considerably higher (i.e. less deprived) than average for MPANs with EV and slightly higher than average for PV across profile classes. For Heat Pumps results varied by profile class. Note that the IMD is 'Index of Multiple Deprivation'.

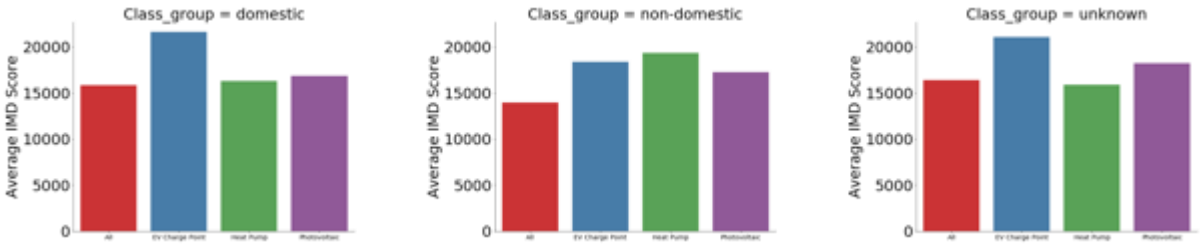


Figure 4-11 – England Class Group

Key

All	
EV Charge Point	
Heat Pump	
Photovoltaic	

The distribution of MPANs with and without LCT has been compared. The deprivation distribution for properties with EV is skewed strongly towards less deprivation, particularly for domestic properties. PV had peaks for low and high deprivation in domestic properties.

Wales Analysis

Average deprivation was considerably higher (i.e. less deprived) for MPANs with EV and HP.

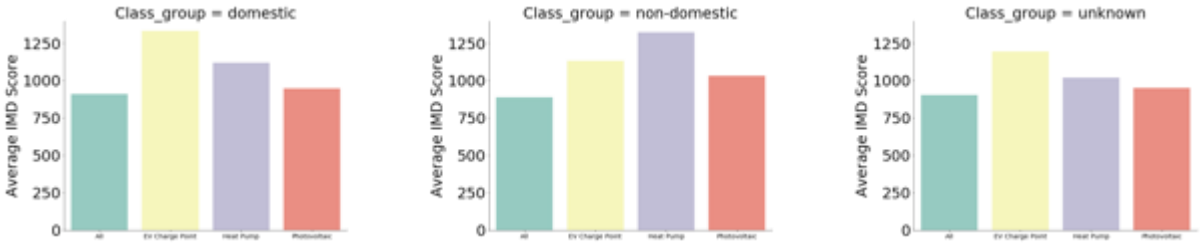


Figure 4-12 - Wales Class Group

Key

All	
EV Charge Point	
Heat Pump	
Photovoltaic	

Making inferences from the Welsh data is more difficult due to smaller sample sizes. The distribution of domestic properties with EV is skewed towards less deprivation, similar to English results.

4.6.2 Consumption Analysis: Introduction

Hypotheses related to consumption changes for MPANs with LCT present have been investigated. Specifically, analysis has been conducted to see what the typical electricity usage is before and after the installation of LCT.

This has been calculated by looking at the change in average daily meter reading prior to installation and comparing it with the change in average daily meter reading after installation.

Consumption Analysis: Results

The analysis of consumption data indicates that a model trained with features related to consumption change can learn to identify instances of PV.

PV Analysis

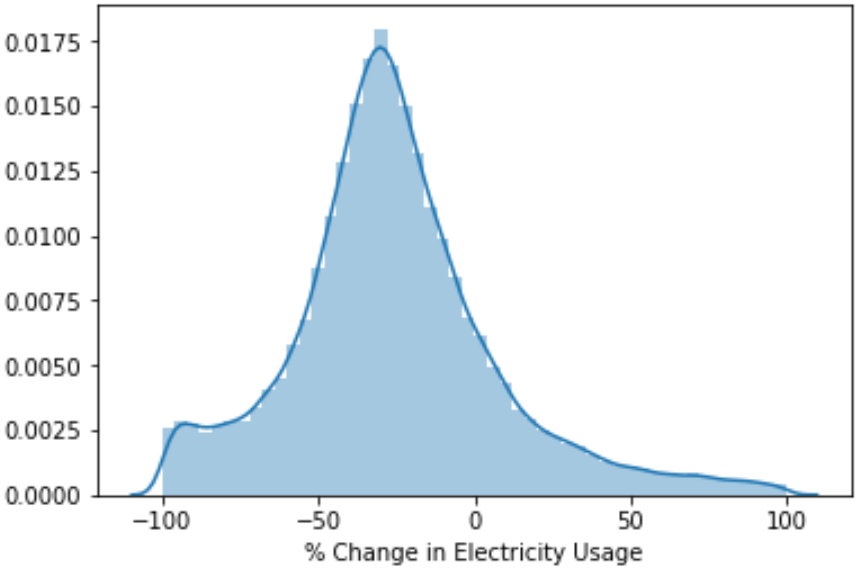


Figure 4-13 – Distribution of percentage change in average consumption per MPAN before and after PV installation

There is an approximate normal distribution centred around the -25% mark (indicating a 25% reduction in electricity usage post-PV installation).

A small number of instances where percentage change is greater than 100% have been excluded in this analysis as they are likely due to data quality issues in the reading data. Additionally, instances close to -100% are likely due to anomalies in the readings data (see data quality issues section for more details).

EV Analysis

The analysis indicates that a model trained with features related to consumption change can learn to identify instances of EV.

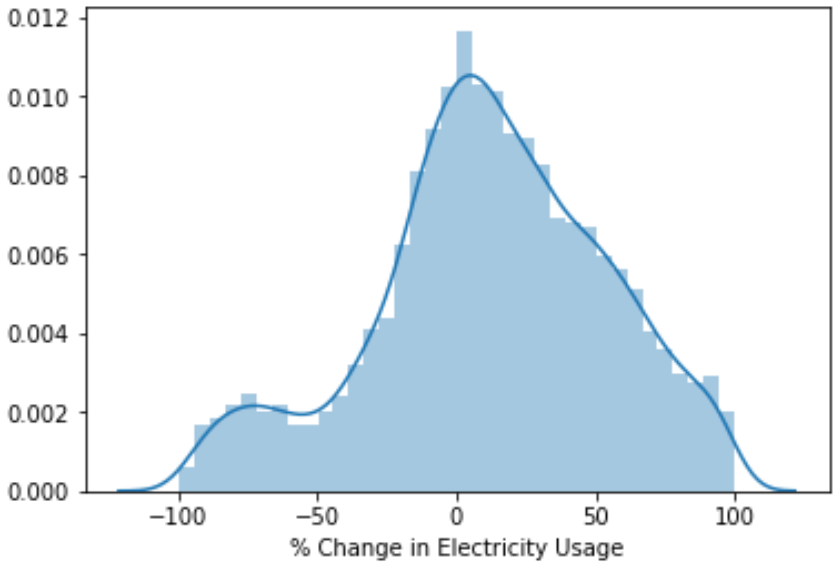


Figure 4-14 – Distribution of percentage change in average consumption per MPAN before and after EV installation

The positive skew of the distribution shows that, in the majority of cases, there is an increase in electricity usage after an EV Charge Point has been installed.

Instances close to -100% are likely due to anomalies in the readings data.

The distribution for EV change is a less defined normal curve than it is for the equivalent PV. This is partly due to a smaller sample set of EV instances and EV Charge Points being introduced more recently, meaning we have fewer post-installation readings.

Heat Pumps

Due to the small number of heat pumps and no consistent patterns of consumption change being visible it is unlikely that a model trained to identify them using consumption patterns would perform well.

The number of heat pumps is much smaller hence it is harder to make inferences from the distribution. The distribution shows a skew towards a reduction in consumption however the number of instances where consumption is -100% suggests data quality issues.

Decision: Remove heat pump analysis from the overall modelling for the LCT detection programme. Consumption Analysis: Data Quality Issues

4.6.3 Other Issues in Sprint 3

Outlying Data

When running analysis on the entire dataset and MPANs with LCT, unusual results were encountered (for example high negative values when analysing consumption between readings).

Investigation showed one cause to be occasional instances of extremely high meter reading values. Typically, these are one-off bad readings for an MPAN with the rest of that MPAN's readings looking reasonable.

A resolution to the issue is to have a “cut-off” value for readings (i.e. if a reading is above a certain high value then those records are removed).

Meter Resets to Zero

Another issue identified is the phenomenon of the meter ‘resetting to zero’ when it passes 99999.

In this instance, the effect is clearly down to the resetting of the meter’s dial and not due to PV, and so instances affected by this will need to be corrected or removed before modelling. Alternatively, the model could be amended to automatically adjust for this specific scenario, based on the meter technical details (number of dials).

Technical Issues in Sprint 3

The analysis of the scale of the issue was delayed by the IBM team experiencing technical issues - the SPARK environment used for analysis on big data sets, such as readings data, was running slowly. A mitigation plan of migrating the existing data and code to a different environment (Analytics Engine) was instigated over the Christmas break, while IBM engineers worked on debugging the speed issue. The issue was discovered to be caused by an upgrade of a different component within Watson Studio which affected some SPARK environments, of which ours was one. The issue has now been fixed, and the analysis is being run in SPARK as originally planned.

4.6.4 Initial Feature Engineering

An initial list of features has been created. Features are inputs into the machine learning model to help train the model which the model will use to understand the key identifiers for LCTs. Once trained, the model will learn to split and combine these features in various ways in order to classify periods of consumption where LCT is present.

When training the model – different combinations and transformations of features will be experimented with to see which yield the best results.

As modelling progresses features are likely to be refined or new features developed.

The following list of features to be developed has been identified:

Feature	Purpose
% change in average daily consumption compared to the same period last year for the same MPAN	Consumption change
% change in average daily consumption since last reading for the same MPAN	Consumption change
Flag for a decrease in consumption since last reading for the same MPAN	Consumption change
Seasonal flag (i.e. a flag if all or part of the period between readings occurs in a given season)	Seasonality
Seasonal percent (i.e. the percent of the period between readings that occurs in each season)	Seasonality
Seasonal flag for median date (i.e. the season the median date for the periods between readings occurs at)	Seasonality
Month flag (i.e. a flag if all or part of the period between readings that occurs in a given month)	Seasonality
Month percent (i.e. the percent of the period between readings that occurs in each month)	Seasonality
Month flag for median date (i.e. the month the median date occurs at for the periods between readings)	Seasonality
Average year on year consumption change for a given year-month across all households in the dataset for the group being modelled (i.e. domestic customers)	Trend

Table 4-6 Features

Consumption features are included as ways to identify changes in consumption patterns, so the model can learn the differences in changes for periods with or without LCT. Each of these consumption changes and seasonality changes will be built per demographic and meter type to capture demographic differences.

Seasonality features are included so that the model can learn the difference between seasonal variation in consumption as opposed to changes due to LCT installation.

A trend feature has been included to so that the model can learn the difference between general changes in consumption (for example, a warm winter leading to less consumption) as opposed to changes due to LCT installation.

Initial Models for Analysis

From analysis of the free text data an additional 1,924 MPANs with PV, 37 MPANs with EV, and 4 MPANs with heat pumps were identified. This equates to an increase of 1.58%, 0.32%, and 0.33% of the number of known instances of PV, EV, and heat pumps respectively.

In this Sprint initial models for categorizing unstructured data have been created.

The unstructured dataset is made up of 8.8 million comments from engineers making site visits, the MPAN the comment relates to, and the date the comment was made. These site

visits are usually in response to a customer query, or to investigate some unusual electrical or meter activity. In general, comments are short, consisting of fewer than five words.

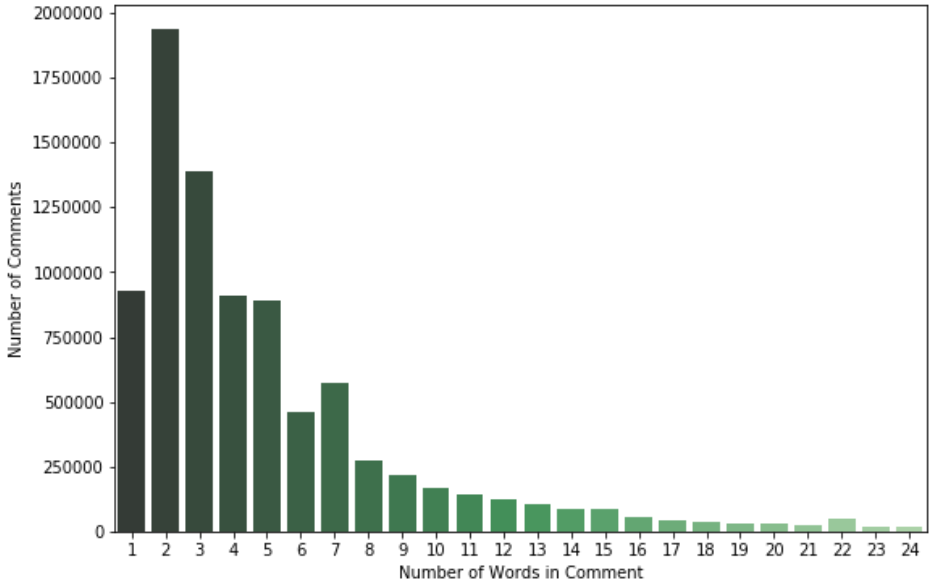


Chart 4-1 - Number of words in comments in unstructured data

82.5% of comments contained 7 words or fewer, with 60% of comments containing 4 words or fewer.

The most common comments were investigated:

Comment	# of times present	% of all comments
no access	387,518	4.38%
<blank>	338,927	3.83%
k	320,652	3.63%
standard service standard service levels	294,845	3.33%
white door	156,424	1.77%
nrq	142,644	1.61%
no answer	141,066	1.60%
unable to obtain a remote meter reading	117,175	1.33%
no more info	103,716	1.17%
brown door	49,481	0.56%

Table 4-7 – Ten most common comments in the unstructured data

None of them give any indication of LCT so are not considered in the following analysis.

Keyword Search

A keyword search was performed on the unstructured data to flag comments that contained certain words or phrases that are attributed to different types of LCT. Once false positives were accounted for, newly discovered instances of LCT were recognised.

The remaining 7.3 million comments were then subject to a keyword search. The list of keywords was provided by ElectraLink and was considered to be close to exhaustive, with any other words likely to either be too ambiguous or too rare.

Some keywords recommended by ElectraLink were seen to produce only false positive results, such as 'pev' causing the flagging of comments including mis-spellings 'pevious' and 'pevgeot'. Additionally, 'ev' and 'hp' produced hundreds of false positives because of either mis-spellings (e.g. 'hpuse') or present at the start or end of words (e.g. 'every'). The nature of the content of the comments means that some keywords, particularly any such acronyms, are susceptible to being mistakenly flagged. For this reason, keywords which were recognised to cause only false positives were removed prior to the search being run.

The keywords were split into three categories: EV, PV, and Heat Pump.

Category	Keywords
EV	'charge point', 'electric vehicle', 'charging point', 'hybrid car', 'range extender', 'tesla', 'model x', 'model s', 'i3 ', 'i8 ', 'outlander', 'car plugged in'
PV	'voltaic', 'solar', 'pv ', ' pv'
Heat Pump	'heat pump', 'ground source'

Table 4-8 – Keywords for each LCT category

Each comment was checked to see if it contained any of the keywords, however, the nature of some of the words leads to some ambiguity: 'charge' can mean electric charge or a financial charge to a customer, 'panels' can be solar panels or door panels, 'pv' can be shorthand for photovoltaic but is also present in pvc plastic. This meant that initial keyword search results contained a lot of false positive results.

The following phrases were therefore identified as causing such false positives from initial searches and were removed from any comments in which they appeared: 'excess annual charge', 'debt charge', 'standing charge', 'upv', 'wpv', 'pvc', 'glass panels' and 'door'.

The keyword search output the following results:

Category	# of comments	# of MPANs
EV	64	39
PV	3,884	3,042
Heat Pump	5	4

Table 4-9 – Number of comments flagged as containing any of the keywords

When comparing MPANs that have been flagged by the keyword search algorithm with MPANs that are in the dataset of known LCTs, we see the following comparison:

Category	# of MPANs in tagged by keyword search	# of MPANs tagged by keyword search that are not in known dataset
EV	39	37
PV	3,042	1,924
Heat Pumps	4	4

Table 4-10 – Comparison of whether MPANs flagged by keyword search are in known data or not

A Watson Natural Language Classifier was built and linked to the Watson Studio project. The model was trained on the 3884 comments classified by the keyword search, with the negative instances being a random sample (4000) of those comments not tagged by the keyword search. This classifier will take comments not previously recognised as indicating the presence of LCT and seek to recognise if LCT is indicated. Due to the long-expected runtime of running all comments through the model (>7 hours), a sample of 200,000 comments were passed through and the results checked.

The model outputs the text it has checked, alongside the likelihood (or confidence) of the text containing indications of PV, EV, or Heat Pumps. Alongside this it gives a likelihood score for whether the text contains evidence of there definitively being no PV, and a likelihood score of the text containing no evidence one way or the other.

Only 129 of the 200,000 comments were considered by the Watson Classifier to have more than a 50% chance of containing evidence of PV. Human analysis of these most promising comments showed that all were false positives (text that was classified as indicating the presence of PV but actually not doing so). The low number is likely due to the fact that not many (if any) of the comments not already classified by the keyword search actually contain evidence of PV. The widespread presence of false positives is probably caused by the training data items containing just a few words, meaning it assigns too much importance to what it initially perceives to be positive indicators.

For EV and Heat Pumps, the results and the reasons for the results are very similar to those for PV. However, this is also compounded by the relatively small training set that could be used for the model (61 and 5 labelled instances respectively). Such small training sets makes it difficult for the model to recognise any patterns, thus making future classification less likely to be accurate.

An overriding reason for the low numbers of positively classified comments is also caused by the success of the keyword search which found an extra 1,961 instances of LCT that were not previously recognised as such. Thus, any further comments that do show evidence of LCT will be more difficult to find and will likely be of a smaller number than those flagged by the

keyword search. To conclude, the keyword search succeeded for at least the majority of the comments with positive evidence.

4.7 Sprint 4 Summary

4.7.1 Final Model: Summary

Sprint 4 is the final sprint of the LCT detection project. The LCT detection project has proved the concept that utilising ElectraLink's DTS dataset and IBM's data analytics capabilities, WPD can identify previously unknown LCTs on their network.

The project will deliver to WPD two PoC models (one model based on analysis of free text comments and one model analysing consumption changes) that can identify LCTs in WPD's network and the supporting code that is required to support the delivery of these models.

Whilst this project has proved the concept of these modelling techniques, the project has also identified a number of opportunities to improve the accuracy of the modelling output and the model's ability to identify LCTs that we would recommend WPD take forward in future.

Summary of Sprint 4 progress:

- Code: The project will deliver Code to WPD that 'normalises' the values within domestic EAC records to take account for the way the data is used in settlement processes to ensure that ElectraLink's DTS data can be structured in a way that can be understood (for example, taking account for inconsistent/incorrect readings, a new meter being installed and when the readings return to zero);
- PoC - Model 1: The project will deliver Code to WPD that can be used to spot EV, heat pumps and PV from freeform text; there will be a clear recommendation in terms of BAU that field engineers automatically identify LCTs as a requirement of the relevant form that feeds into the free text fields, irrespective of whether there is an issue relating to the LCT;
- PoC - Model 2: A model that can identify the installation of PV through consumption changes has been created.

Prior to Sprint 4, the intention was to create a model that could identify PV and EV installations in the WPD area. During Sprint 4, it became clear that it is not possible for the model to identify all PV and EV installations and the reasons are, as follows:

- Data 'noise': As expected, there are a number of factors that affect consumption in a household other than the existence of LCT which creates 'noise' in the data, i.e. unexplained changes in consumption. With only one factor 'known' to the model (the presence of LCT) to understand consumption changes, this data noise means that the model can experience difficulty identifying correctly whether the LCT is present or the change in consumptions is due to other reasons.
    a. In project mitigation: In response to this, the modelling was targeted to focus on MPANs with more consistent consumption changes.

b. Post project mitigation recommendation 1: By adding more datasets that can explain changes in consumption, i.e. weather data, we can begin to eliminate this 'noise' as the model will learn the nuances of consumption.
c. Post project mitigation recommendation 2: With the introduction of more granular data (such as Smart data), we can begin to understand more accurately the seasonality or time-bound nature of the consumption changes which would support understanding of the potential for these changes to be driven by PV or EV.
d. Post project mitigation recommendation 3: It is possible that the positive dataset for EV could only indicate the presence of a charge point and not an EV; therefore, understanding whether the positive dataset is a complete and accurate illustration of known EV in the WPD area will enable IBM to model the consumption changes more accurately.

- Model: Similar to the data 'noise' issue, lack of a negative data set is a challenge to the modelling as it does not allow the model to learn non-LCT 'normal' consumption changes – the creation of a negative set will eliminate this risk. See table below for impact, mitigation and progress.
    a. In project mitigation: In sprint 2 we agreed to use consumption data for known LCTs prior to the installation date as the negative set. The model has been created on this set. This means that we cannot use other factors affecting those MPANs (e.g. demographics). The reason for this is that the model will look at all the factors that lead an MPAN to have an LCT to determine the likelihood of having an LCT; for example, the model will look at a consumption increase that is consistent with the presence of an EV and, if the MPAN is in an affluent area, there is a high confidence the MPAN has an EV. If the consumption jump was for an MPAN in a less affluent area, given that less affluent MPANs are unlikely (due to the analysis) to have an EV, the model would take that into account and the confidence level in the MPAN having an EV would be lowered. This type of demographic analysis is only possible if we had a dataset of known no-LCT MPANs for the model to 'prove' that LCTs are less likely in a less affluent area. The data that we have only highlights the demographics of the LCTs that are known; therefore, if we train the model using the consumption patterns of MPANs with an LCT before they have an LCT to understand the consumption patterns of 'before' and 'after', we cannot also use the demographics because the model would then assume that these factors determine that the LCT is not present. By using the known MPAN's demographics in the no-LCT dataset, the model would believe that the presence of a MPAN in an 'affluent' area would be a defining factor for them not having an LCT, when the opposite is true.
    b. Post project mitigation recommendation 3: Secure a negative dataset that outlines the households that do not have an LCT installed.

Project Findings:
The project was able to identify 13,967 (5,863 EV and 8,104 PV) previously unknown LCTs in WPD's network.

The following section outlines how we achieved this understanding.

4.7.2 Sprint 4 Outputs

Additional Data Cleaning

The consumption model uses features derived from energy consumption for a given MPAN. Analysis undertaken in Sprint 3 identified numerous anomalies in the consumption data. In particular, unusually high meter readings, unusually low meter readings or MPANs for which the readings appear to "reset" to zero. These data quality issues discovered could lead to errors in the model so data cleaning was required prior to modelling.

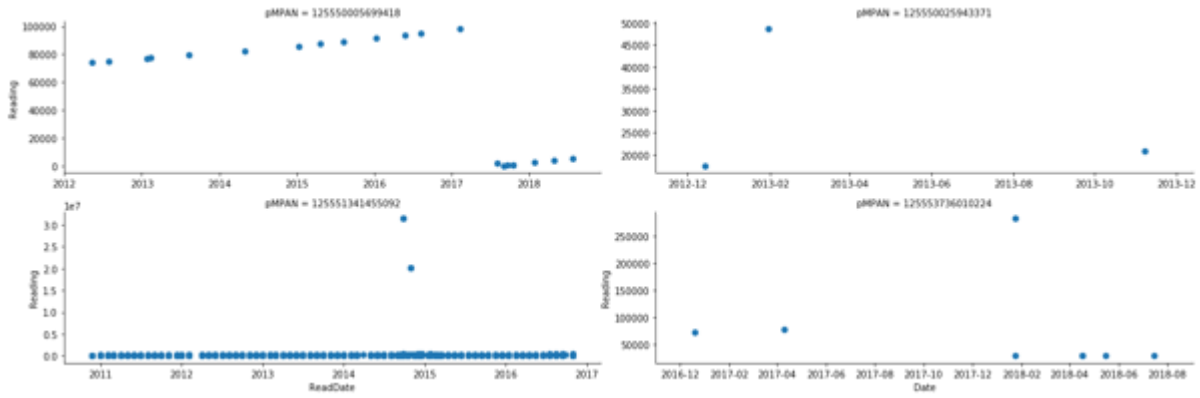


Chart 4-2 - Some examples of MPANs with different anomalous behaviours including spikes and drops in readings, as well as meters resetting.

The following data cleaning tasks were completed after some features were constructed measuring changes in consumption:

- Readings with a Read Type of 'W' (indicating the reading was withdrawn after being taken) were removed as their values could not be trusted.
- Remove readings taken on the same day as the previous reading.
- Remove readings where there was a drop of less than 1kWh from the previous reading (mostly due to unusual readings with too many decimal places).
- Remove solitary drops in readings – readings where there has been a drop by at least 50% and immediately jumps back up to the previous level.
- Remove solitary jumps in readings –readings where there is an increase of at least 50% followed by the readings dropping to their previous level.
- Remove MPANs where there is a drop of at least 95% that does not come back up to the previous level immediately after. Most often this was because of meters looping through to zero, but amending readings to take this into account is made difficult by the different lengths of meters and that meters seem to loop at otherwise arbitrary values.

4.7.3 Additional Analysis

Following additional data cleaning, analysis was done to guide the modelling approach.

The analysis in Sprint 3 identified that there were sporadic changes in consumption for EV and sometimes PV. The project hypothesised that this could be driven by the number or frequency of meter readings performed following the installation of the LCT. For example, if only 1 meter read was captured following LCT installation, it is not guaranteed that the total true impact of the LCT installation will be captured in this read. Therefore, additional analysis into the reading was performed and the questions investigated are as follows:

1. What is the number of readings before and after installation of LCT (EV, PV)?
2. What is the average time between readings before and after installation?
3. What is the frequency of readings for different meter types?

An investigation into the number of readings pre- and post-LCT installation was undertaken to examine the suitability of the data for modelling. To train a machine learning model, a sufficient number of readings pre- and post-LCT installation is needed to determine a change in consumption behaviour. An overall count of readings pre- and post-LCT do not indicate any issues for modelling in this regard.

Read_GEN_TYPE	count	Read_LCT_TYPE	count
None	674765	None	175567
Photovoltaic	2729604	EV Charge Point	127255

Table 4-11 - Readings Pre and Post LCT Installation

Similarly, if we break down the numbers of readings by class group, we see the proportions of readings stay fairly consistent (excluding non-domestic for PV, which appears to show a different trend to domestic and unkown).

Read_GEN_TYPE	Class_group	count	Read_LCT_TYPE	Class_group	count
None	unknown	25657	None	unknown	4944
None	domestic	589510	None	domestic	165431
Photovoltaic	domestic	2373524	EV Charge Point	non-domestic	2788
Photovoltaic	non-domestic	182042	EV Charge Point	unknown	4752
Photovoltaic	unknown	174038	EV Charge Point	domestic	119715
None	non-domestic	59598	None	non-domestic	5192

Table 4-12 - Readings Pre and Post LCT Installation by Class

The average number of days between readings pre- and post-LCT installation was investigated to look for any major differences in the frequency of readings. The results showed that although there was a small increase in the average frequency of readings after

installation for both PV and EV, the difference does not appear to large enough to cause an issue when modelling.

Read_GEN_TYPE	avg(DaysSinceLastReading)		Read_LCT_TYPE	avg(DaysSinceLastReading)
None	75.47918426652137		None	77.90634889362279
Photovoltaic	65.97071790807759		EV Charge Point	67.58155025099225

Table 4-13 – Reading Intervals

Finally, the average frequency of readings for the different meter types was considered. Very long periods between readings has the potential to impact modelling negatively, as there are fewer observations on which to train the model and make predictions. Although we cannot be sure whether the average frequency of readings of the Non-Smart or Prepayment meters will cause issues when modelling until we build and train the model, the significantly lower frequency of the readings may potentially cause issues. The model uses an engineered feature (daily consumption) which is an interpolation of the non-uniformly distributed consumption data points. As a part of the validation and elaboration of the model in phase 2 of the project – we can assess meter read frequency as a confidence factor in the prediction of LCTs.

MType_group	avg(DaysSinceLastReading)
Smart	37.02326341921678
Others	35.188993267119 83
Prepayment	63.17931551869645
Non-Smart	98.52526720696177

Table 4-14 – Reading Intervals by Meter Type

Once we began modelling, it became clear the variability in the frequency of readings between different meter types was causing an issue in modelling. As the readings were so sporadic, there were too few periods in which the model could predict LCT. Considering this, we chose to narrow our focus to modelling domestic MPANs with smart meters.

4.7.4 Initial Modelling Approach

The initial approach taken when modelling was to build relevant features on the entire dataset and then train a model to use those features to predict periods with LCT.

The features were all built around consumption, as other features around demographics could not be used due to the absence of a negative dataset. Furthermore, the absence of the negative dataset meant we were unable to build a predictive model to predict whether an individual MPAN (controlled for meter, register and tenant) has LCT, but instead predict whether *particular readings* for a given MPAN indicate LCT. This is because without knowing which MPANs do not have LCT, we cannot train a model to find the difference in consumption patterns or demographic characteristics between MPANs with LCT and MPANs without LCT.

Project Learning:
The project recommends that additional work is undertaken to develop a negative dataset to understand consumption changes in more detail.

The features we initially built for experimentation were:

- Average daily consumption
- Change in average daily consumption to the previous reading
- Month
- Year
- Season
- Change in average daily consumption to the previous year

The feature selected to build the model was the change in average daily consumption compared to the previous year. This was selected as:

a) using average daily consumption instead of consumption between readings controls for differences in the frequency of readings, and
b) comparing consumption to the previous year controls for the seasonality effects on energy consumption.

The derived target variable for the model is whether the reading is within the 12 months after LCT installation. This is therefore when a reading's 'Read_GEN_Type' or 'Read_LCT_Type' indicate LCT was present for the reading, but for the same period in the previous year the 'Read_GEN_Type' or 'Read_LCT_Type' show LCT was not present. Therefore, for each MPAN with known LCT, there will be 12 months of readings as the target variable the model is trying to predict.

The model chosen was a binary random forest classifier. Random forest classifiers utilise 'decision trees' – collated together into 'random forests'. Decision trees map datasets to their outputs with a set of questions with the model flowing through these binary questions to understand whether the MPAN has the LCT characteristics, for example: A set of the questions could be 'has there been a drop in consumption?' with a set of binary answers 'Yes' nda 'No'. If yes, the next question could be 'is this consumption change consistent across the subsequent readings?' with another set of binary answers of 'Yes' and 'No' and this will go on until the model has confidence in the consumption change being linked to LCT installation. The intention is to collate a number of these decision trees, with a number of different questions around the MPAN to minimize the number of overarching assumptions on the dataset and minimize the risk of model overfit.

The base of the decision tree for this model is the decision that splits the dataset, using the most defining feature (has there been a consumption change). The following questions will then try to identify the characteristics of the consumption change and whether it matches the installation of an LCT.

Random forests is an ensemble classifier which uses many decision tree models to predict the result. A different subset of training data is selected, with replacement to train each tree. The trees are being trained on subsets which are being selected at random, hence random forests. Class assignment is made by the number of votes from all the trees.

4.7.5 Initial Model

Initially, each model was built using all the MPANs with the known LCT, with the aim of modelling the impact of LCT installation on each MPAN's consumption of energy.

Building the model on all LCTs proved unsuccessful because the inconsistency in the frequency of readings between different meter types and between non-domestic and domestic MPANs meant that a consistent view of consumption changes was not able to be identified. Therefore, the data for modelling was restricted to smart meters in domestic MPANs.

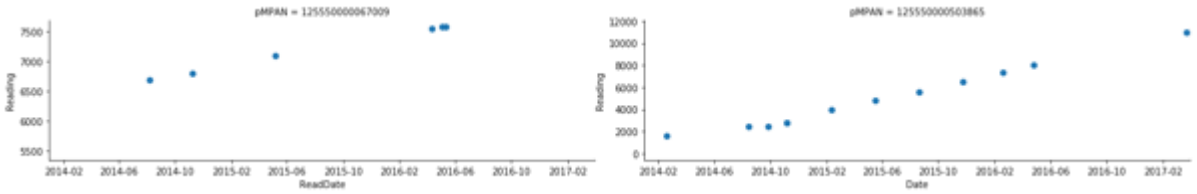


Chart 4-3 - Example of two non-smart meters showing very infrequent readings

However, even after filtering the data to use only the smart meters in domestic MPANs, a predictive model to understand the impact of LCT installation still could not be built as the signals for LCT were lost in the noise of seemingly inconsistent behavioural changes in the MPAN's consumption. Examples of this can be found below:

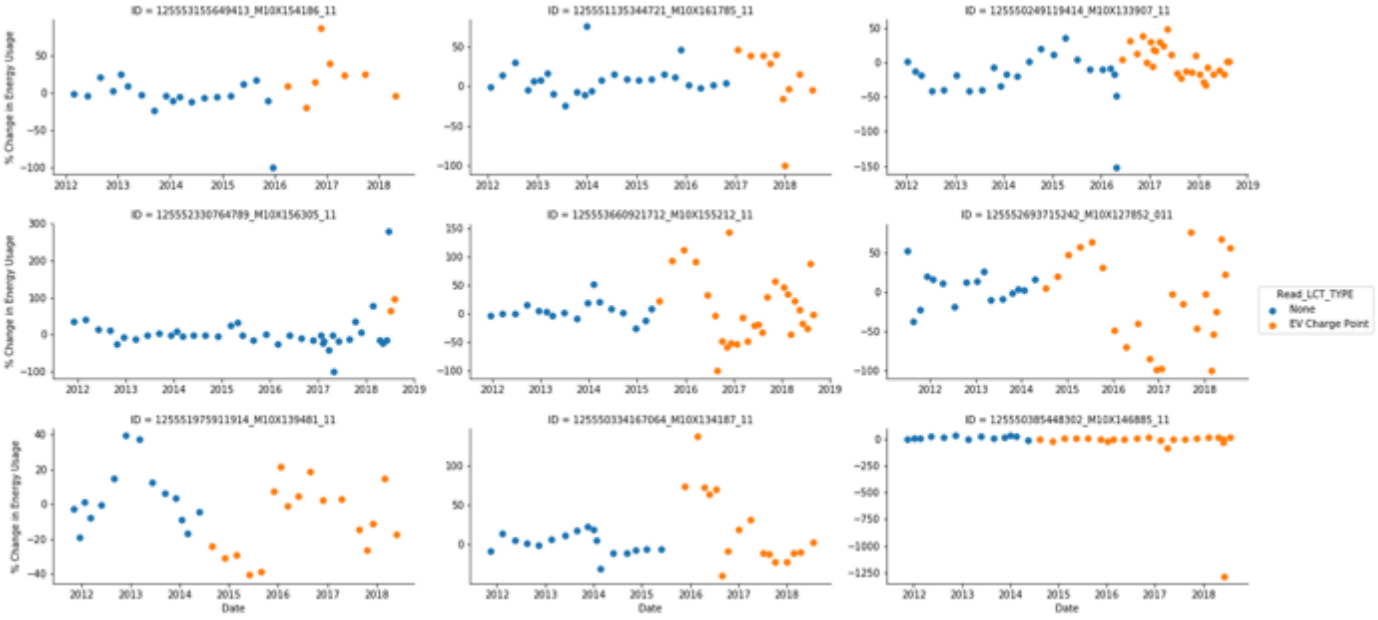


Chart 4-4: Example of a random selection of smart meters in domestic MPANs, showing very variable changes in daily energy usage compared to the same period in the previous year.

4.7.6 Devising a new Modelling Approach

Smoothing data

During previous analysis, it was apparent that there were MPANs with big spikes or drops in year-on-year change in consumption, mostly caused by short periods of high or low daily consumption that were likely caused by other factors. These spikes and drops distorted our attempts to identify a strong signal. To combat this, a rolling average of the year-on-year consumption changes from the last 90 days was constructed, smoothing out the spikes and drops present in some MPANs' plots. The result is fewer extreme values, helping to more easily identify MPANs with little noise.

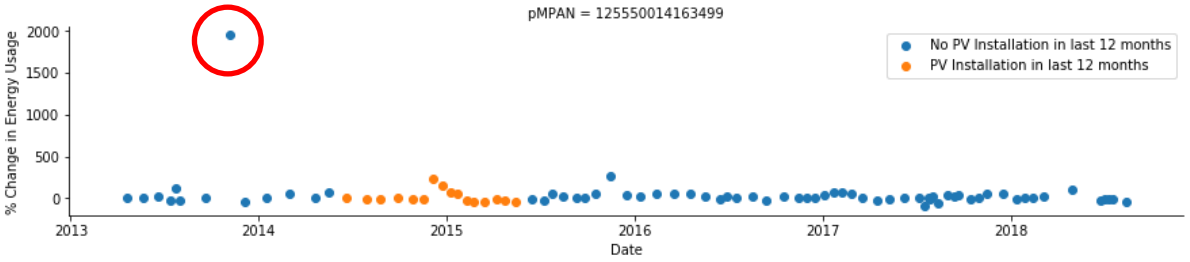


Chart 4-5 - An example of a spike in consumption change which affects measures of noise

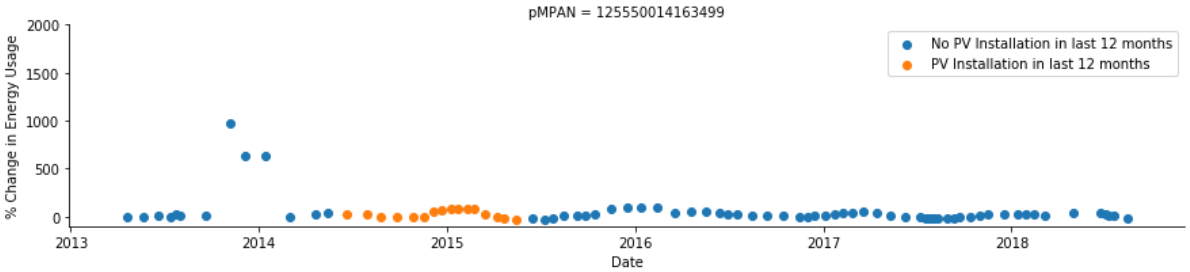


Chart 4-6 - An example of the rolling average of consumption change

Identifying MPANs with the Strongest Signal

Despite the benefits of the rolling average, the difficulties caused by the noise in an individual MPAN's energy consumption changes were still apparent, so the decision was made to identify the MPANs with the least noise and the strongest signal (consumption increase after EV Charge Point installation and consumption decrease after PV installation). The rationale for this was that a model could be trained to recognise these signals and would be able to identify MPANs with similarly strong signals in the unknown dataset.

The plot of an ideal signal would show consistent energy usage prior to LCT installation, a noticeable year-on-year change in the year after installation, followed by a return to little year-on-year change once the LCT has been present for more than 12 months.

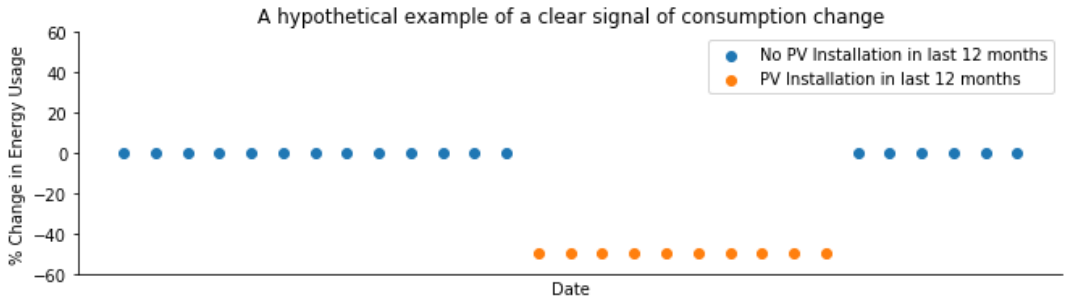


Chart 4-7 - A hypothetical example of a clear signal of consumption change

Identifying the MPANs with the strongest signal and least noise was achieved by considering the following characteristics of each MPAN:

a) The average daily consumption in the 12 months prior to LCT installation relative to the 12 months after LCT installation. The larger the difference between these, the stronger the signal for the model to identify;
b) The standard deviation of the rolling average change in daily consumption for all the readings of an MPAN. A larger standard deviation means more variability in consumption (i.e. more noise).

Each MPAN was ranked according to each of these characteristics, indicating their suitability for being included in building the model. Determining where to set the thresholds for these metrics is a matter of experimentation and needs to balance conflicting factors. Stricter thresholds will reduce the modelling subset to only include those with the strongest signals, meaning the resulting model will identify only MPANs with similarly strong signals. However, more relaxed thresholds will increase the amount of MPANs in the modelling subset with weaker signals, meaning more LCT predictions in the resultant model but with the associated risk of more of these predictions being false positives.

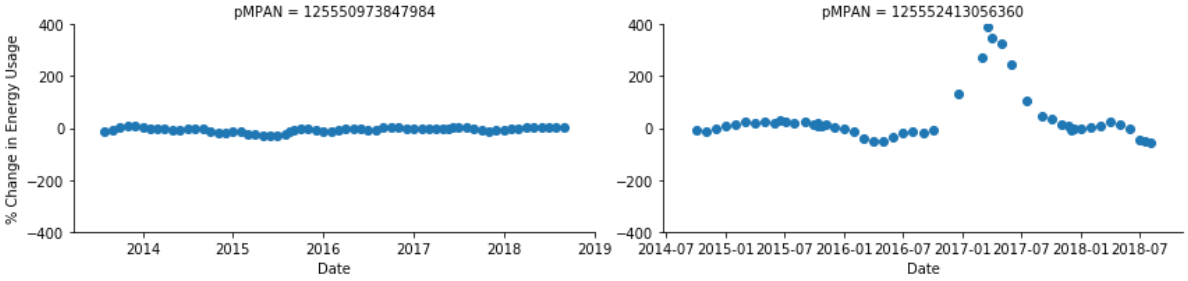


Chart 4-8 - An example of low standard deviation and high standard deviation for two MPANs' changes in energy consumption

Interpreting model outputs

As we model on a reading level, each MPAN can return multiple predictions of LCT. For the output of the model to be at an MPAN level, these results need to be aggregated. Despite the smoothing of the rolling average calculations, some MPANs still demonstrated solitary spikes or drops in consumption change meaning that they often returned positive predictions for only a single reading.

A better indication for the recent installation of LCT is instead something resembling a year-long change in year-on-year consumption behaviour. However, this is complicated by the variety that MPANs have in number of readings per year, so a simple rule of "any MPAN that returns *n* positive predictions can be considered to show evidence of LCT" is not strong enough. Instead, a ratio of the number of predictions of LCT to the average number of readings per year was constructed for each MPAN, and this ratio can be used to decide what is reasonable to consider as an MPAN-level positive prediction. This will ensure that the MPAN consistently shows a consumption change *over time.*

$$rolling\ average = \sum_{i=0}^{i=90}(\Delta\ Consumption) \Big/ number\ of\ readings$$

Where i represents the days to look back over.

A further indication of LCT installation on an MPAN level is the time between the first and last predicted period of LCT. As we are measuring consumption change compared to consumption in the previous year, we are predicting the periods in the 12 months after LCT installation. Therefore, the closer the time between the first and last predicted period of LCT is to 12 months, the more likely this MPAN has had LCT installed.

4.7.7 Modelling Approach

Formalised approach

1. Import cleaned dataset.
2. Filter for smart meter types and domestic MPANs.
3. Split dataset into known LCTs and unlabelled MPANs.
4. Rank known LCTs by the size of the change in average daily consumption since installing LCT and the standard deviation of the change in average daily consumption. Take the MPANs with the highest rankings, and therefore the MPANs with the strongest consistent signal.
5. Build and train a classification model using these MPANs to predict readings in which LCT has been installed in the previous 12 months.
6. Apply the model to the unlabelled MPANs.
7. Aggregate the predictions up to an MPAN level.
8. Select MPANs that show strong signs of LCT.

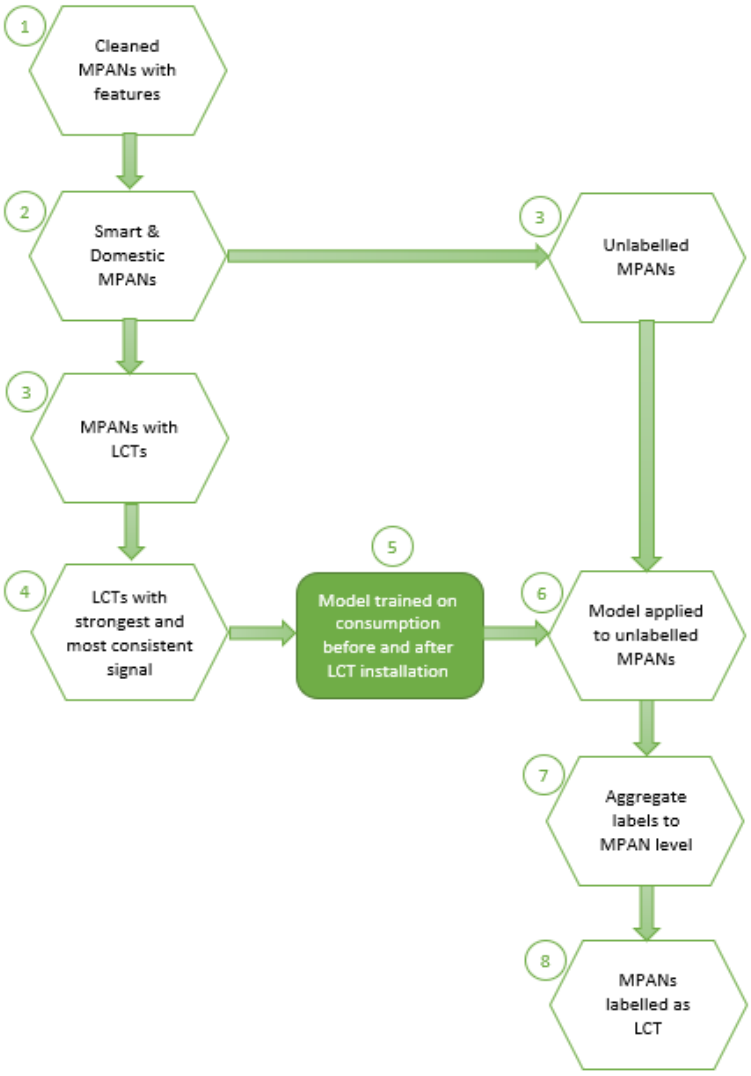


Figure 4-15 - Overview of our final formalised approach

As outlined in Sprint 2, the cleaned dataset excludes any MPANs that have been flagged for Change of Tenancy as there is no guarantee that the consumption change does not relate to the change in the household occupiers.

Rules for selecting MPANs and number of MPANs this selected for test/train EV/PV

MPANs known to have LCT (filtered down to just domestic and smart meter types) are ranked by:

1. Change in average daily consumption in the year before installing LCT compared to the year after;
2. Standard deviation of change in average daily consumption.

Once these MPANs are ranked, a proportion of the MPANs with the largest change in consumption after installing LCT and the lowest standard deviation of change in average daily consumption are used to train the model.

Stage	**MPANs** (Controlled for meter, register and tenant)
Smart, Domestic, EVs	2,192
Insufficient readings to calculate standard deviation	1,677
Insufficient readings – no readings in the year pre- or post-installation	1,106
Insufficient readings - 1 reading prior to installation	774
Less than 5 readings before or after installation	291
Top 150 rank for change and top 150 rank for standard deviation	65

Table 4-15 - EV MPANs for Modelling

Stage	**MPANs** (Controlled for meter, register and tenant)
Smart, Domestic, PVs	21,550
Insufficient readings to calculate standard deviation	17,223
Insufficient readings – no readings in the year pre- or post-installation	6,139
Insufficient readings - 1 reading prior to installation	3,321
Less than 5 readings before or after installation	1,333
Top 600 rank for change and top 900 rank for standard deviation	416
Top 300 rank for change and top 300 rank for standard deviation	55

Table 4-16 - PV MPANs for Modelling

Model Test Results

For the consumption models we are using a random forest with 200 trees. The known LCTs are split into a training and test set using an 80% - 20% ratio. This means we can build the model and then evaluate the accuracy of the model's ability to predict readings 12 months after the installation of LCT for the LCTs not used to train the model. However, as we select the MPANs that show the strongest and most consistent signal, these figures will be positively skewed increasing the rate of accuracy. The measure of accuracy is the number of periods in the test data that the model correctly predicts. These measures of accuracy however must be viewed with caution, as all these MPANs have been preselected because they show similarly show strong consistent signals.

Predicted		Label	
		No LCT	LCT
	No LCT	388	57
	LCT	51	91

Table 4-17 - EV ROC

EV accuracy = 81.6%

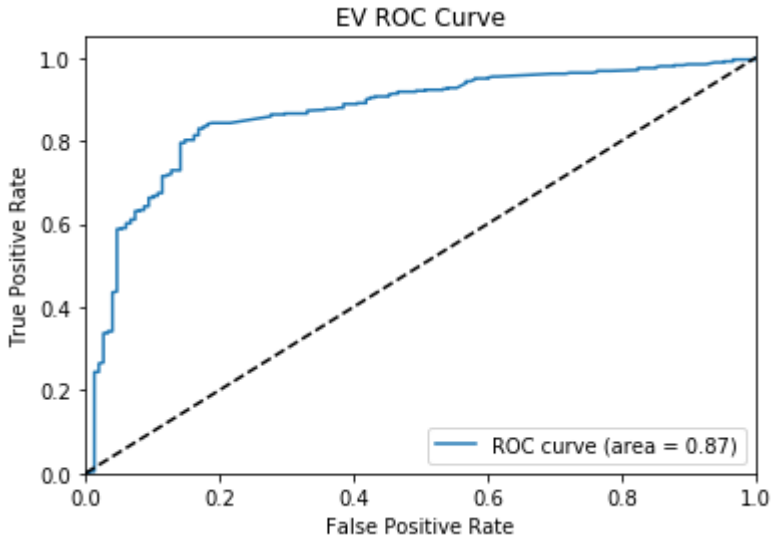


Chart 4-9 - EV ROC

Predicted		Label	
		0	1
	0	396	61
	1	39	73

Table 4-18 - PV ROC

PV accuracy = 82.4%

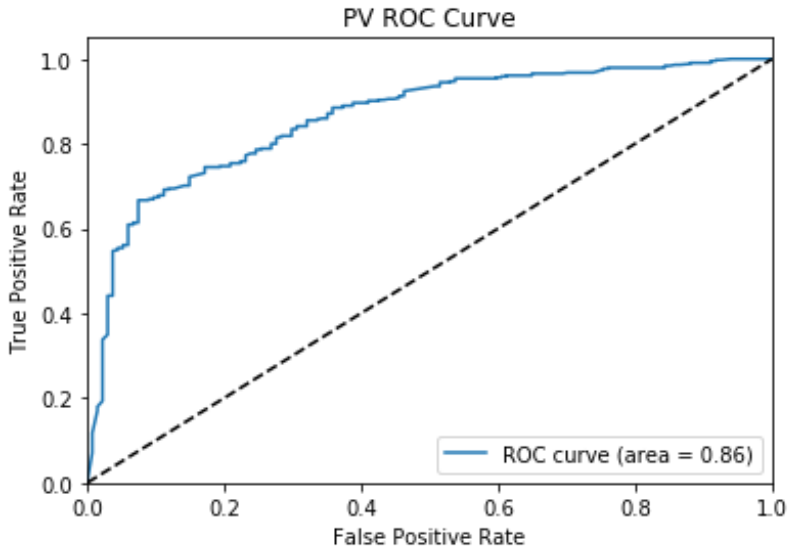


Chart 4-10 - PV ROC

Results of Applying Consumption Model to Unlabelled MPANs

Once the random forest model is trained on the selected known LCTs, we can then apply the model to the unlabelled MPANs.

Metric	No. MPANs
Total unlabelled MPANs	1,049,004
0 periods predicted to have LCT	631,521
1 or more periods predicted to have LCT	417,483

Table 4-19 - EV Results

Metric	No. MPANs
Total unlabelled MPANs	1,026,298
0 periods predicted to have LCT	614,846
1 or more periods predicted to have LCT	411,452

Table 4-20 - PV Results

Once we have all the predictions for each reading, we can aggregate those predictions up to an MPAN level (controlled for meter, register and tenant) and view the MPAN as a whole.

Rules for selection of MPANs as having EV/PV

Two metrics were created to measure whether the aggregated predictions indicate whether an MPAN has LCT. The first was the number of predicted periods of LCT divided by the average number of readings per year. This metrics shows what proportion of an average 12 months of readings has been predicted as having LCT for a given MPAN. This is useful as the perfect prediction will predict a full 12 months of readings as having LCT.

Secondly, the timespan between the first and last predicted period was calculated. Again, the perfect prediction will have a time span of 12 months between the first and last predicted period, and so finding MPANs with a timespan close to a year indicates the MPAN has LCT.

Metric	Thresholds
Predicted Readings / Average No. Readings per Year	0.9 < x < 1.1
Timespan between first and last predicted period (days)	250 < x < 400

Table 4-21 - EV Timespan

Metric	Thresholds
Predicted Readings / Average No. Readings per Year	0.9 < x < 1.1
Timespan between first and last predicted period (days)	250 < x < 400

Table 4-22 - PV Timespan

LCT Type	No. Predicted MPANs
EV Charge Point	5,863
Photovoltaic	8,104

Table 4-23 - Number of MPANs with predictions

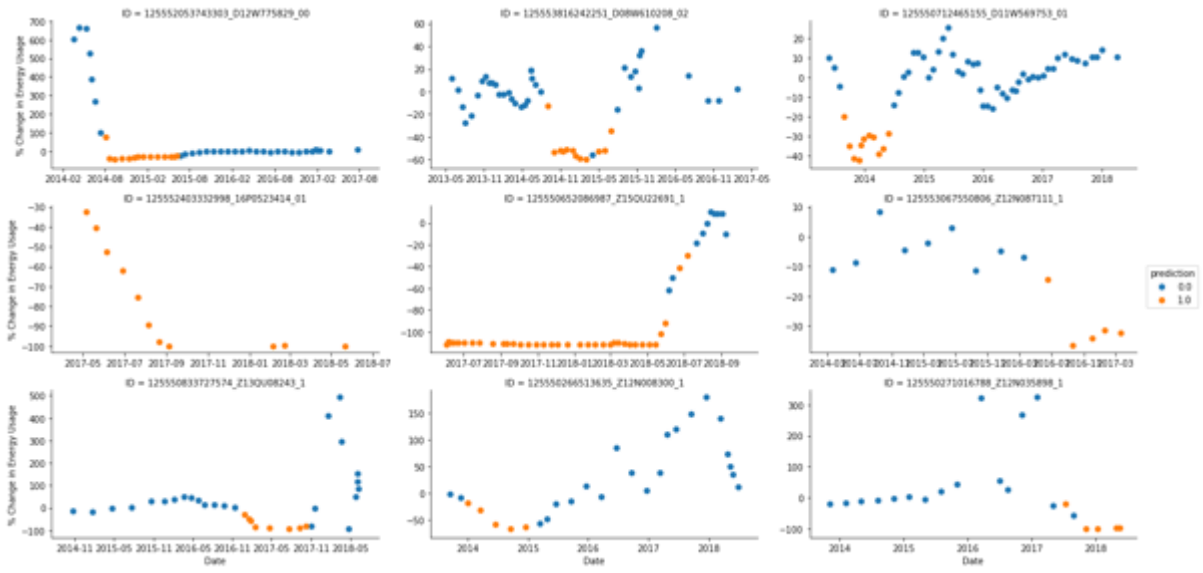


Chart 4-11 - Example Predicted MPANs PV

Plots of a random selection of MPANs which the model predicts as having PV installed and which pass the thresholds detailed above.

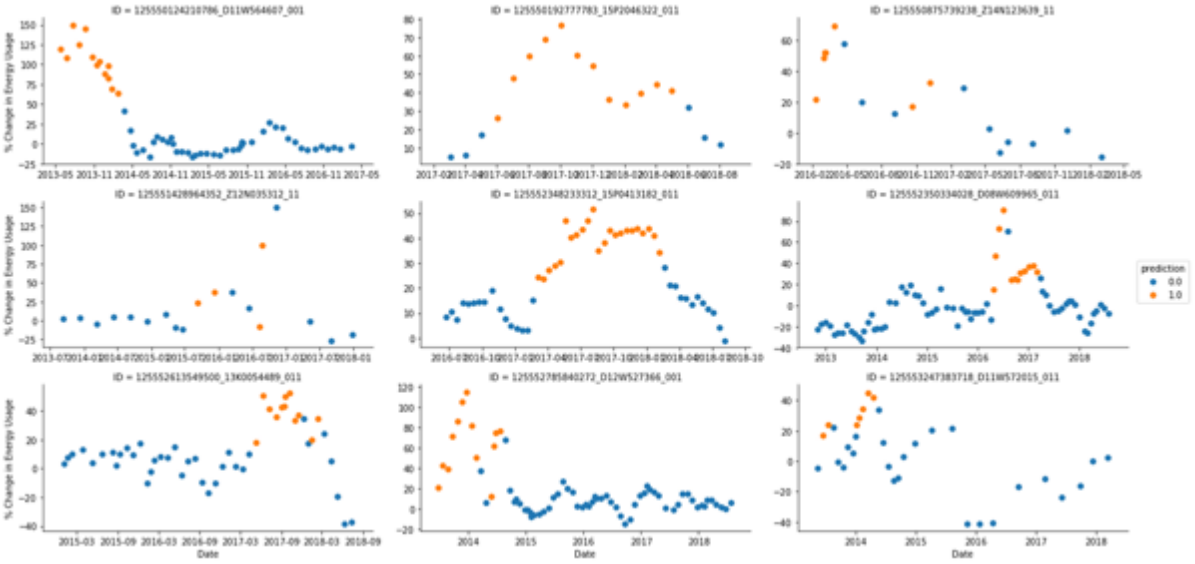


Chart 4-12 - Example Predicted MPANs EV

Plots of a random selection of MPANs which the model predicts as having EV installed and which pass the thresholds detailed above.

In the above graphs, for both LCT types, the orange data points indicate the periods where the model predicts LCT has been installed within the last 12 months. Therefore, a sensible prediction for an MPAN is a 12-month period of orange data points of higher or lower (if EV or PV) than the usual consumption relative to the normal trend.

5 Performance Compared to Original Aims, Objectives and Success Criteria

The LCT Detection project has successfully developed two Proof of Concept models that can spot thousands of previously unknown (unregistered) low carbon technologies connected to the LV network. This project has enabled a potential 13% increase in the visibility of LCTs on the WPD network.

This has been achieved by creating a number of machine learning models that analyse the data from both structured and unstructured data on the DTS dataset. These models can be replicated, with minimal rework, to expand across the rest of the GB DNO network.

The project has achieved all the success criteria.

Success Criteria	Project Success
1. An advanced PoC analytics model that identify LCT on the LV network	The LCT detection project has developed a model that has identified roughly 15,000 potential LCTs on the WPD network that were previously unknown.
2. Introduction of the project to a DNO and energy industry audience at WPD's Balancing Act event in November 2018.	The LCT detection project was discussed at the November Balancing Act event.
3. Presentation of the final report to a DNO and industry audience at WPD's Balancing Act event in May / June 2019.	The LCT detection project will be disseminated at the Balancing Act event.
4. Analysis of data on a number of representative network topologies across WPD's Electricity Service Areas (ESAs).	The LCT detection project covered all WPD areas.
5. Validation – recommendations as to what validation approach WPD should use for each candidate set	The recommendations for validation are detailed in the final report
6. Delivery to WPD of a PoC model a process design document and demonstration dashboard that will identify Low Carbon Technologies (LCTs) to support network planning and investment strategy	The LCT detection project has developed a dashboard to highlight the location of LCTs on WPDs network.

Table 5-1 - Success Criteria

The project has made the following findings against the high-level objectives:

High Level Objective	Project Achievement
* Assuming the model generates a probabilistic output, what level of probability should be used as the cut-off value between the model predicting presence or absence of LCT?	Because of the modelling approach finally adopted – i.e., aggregating potential LCT at a reading level – there is no probability ranking in the model. It is a binary flag LCT or No LCT.
* To what degree can the models produced accurately identify the presence of LCTs not known to WPD and how is the output split between correct identification, false positive and false negative results?	We can only comment on the results run against the test data set and documented in Sprint 4. EV – 82% , PV 82%. Further validation would require an extended validated positive/ negative data set.
* What are the options to validate the presence of LCTs identified by the model? E.g. satellite imagery, comparison to FIT register, site visit, letter to customer etc.	The options for validation will be documented in this report.
* What are the costs for these validation options and how effective are they for the various technologies? Which options are the best value for money for each technology type?	The costs for validation vary based on the scale and scope of validation that WPD will wish to employ. The options for validation documented in this report will be categorised by a HIGH, MEDIUM and LOW cost.
* How is the accuracy of the model affected by the data sets available to it? In particular, given the roll out of smart metering data, how does the availability of data at half hourly resolution affect the accuracy of the model?	The accuracy of the model is greatly improved by the availability of the smart metering data, as the frequency of the reads vastly improved the understanding of consumption changes. Hence the model was focused on smart metering
* Are separate models for domestic and non-domestic customers required?	The modelling found that there is a vast variation between domestic and non-domestic households and different meter types; therefore, we would recommend the creation of a single model for each demographic.
* Can the model be simplified to simple rules of thumb e.g. an increase in EAC of X kWh sustained for 3 consecutive quarters indicates a charging point of capacity Y, a decrease in the August reported EAC of A kWh less than the previous August's value represents a PV installation of 4kW if there is also a decrease in the year on year November EAC of B kWh.	No. The reason we used machine learning is that we cannot use a predetermined equation to understand the presence of LCTs. The rationale for this is that there isn't a standard identifier, such as a standard change in consumption, that can be used to identify an LCT. This is because many factors can affect consumption at a specific MPAN – such as weather and usage – which both limit the opportunities to create simple

High Level Objective	Project Achievement
	equations to understand LCT installations. So rather than programming specific rules, Machine Learning involves using a computer to recognise patterns from examples of LCT installations and turning these patterns into an algorithm (basically a set of rules) that pull out the different ways in which consumption can change with LCT installation. These rules can then be used to make predictions on a wider dataset to see if there are other households follow similar patterns.
* From the WPD records where it is known LCT is installed, can we understand the changes in energy usage from due to LCT installation?	The modelling identified the consumption changes that were driven by the installation of an LCT. This was found in Sprint 3.
* From identification of EVs on the network, can we start to tell the difference between makes of EV, i.e. can we tell the difference between a Nissan LEAF charging at 7kWh and a Tesla Model S charging at 16.5 kWh?(To be discussed)	The granularity of data does not enable the identification of individual EV types. To do this, we believe we would need more granular data (smart and Half Hourly data), plus a 'known' dataset of these cars in installation to train the model and identify the different patterns in consumption.
* Can we achieve 70 – 80% identification of LCTs on WPD's network?	There are many learnings that have come from the project that will improve the coverage of the modelling.

Table 5-2 - Project Objectives

6 Required Modifications to the Planned Approach during the Course of the Project

In all the project went to plan in that we completed four cycles of work with clearly defined objectives in each cycle. There were some technical and other issues encountered which required a change of plan:

Unplanned Team Absence: There were some absences due to illness etc. As the team was small, this issue is difficult to mitigate and the only solution was to use some of the contingency time built into the plan.

Technical Challenges: There were some issues caused by an outage in the Spark environments. The team mitigated this delay by preparing a plan B and working some unplanned days during the Christmas break. Fortunately, the issue was resolved and the team was able to revert to using the Spark environments as planned.

Data Challenges: Two key issues caused a challenge in terms of the data sets:

- No negative set; and
- Limited granularity of data.

Both issues were accommodated in the modelling approach as described in Section 0 Details of Work Carried Out, and in the Sprint Reports.

With both of these issues – the Agile principles of transparency and frequent communication allowed all stakeholders to understand and work with the issues as the occurred.

Removal of heat pumps:
Initial analysis did show promising results for heat pumps with a skew towards a reduction in consumption. However, there were indicators of data quality issues and due to the small number of heat pumps present in the data set it was decided during the project to remove them from the model.

7 Project Costs

Activity	Budget	Actual	Variance	%
WPD Project Management	11,603	8,309	3,294	28
WPD Network Services	370	362	8	2
Project partner costs	302,660	302,660	**0**	**0**
Dissemination Literature	2,150	2,150	0	0
Total	316,783	313,481	3,302	1

Table 7-1 Project Costs

8 Lessons Learnt for Future Projects

Stakeholder Engagement Workshop: The project generated excellent learning in terms of the approach used by IBM for the Design Thinking Workshop in eliciting desired information around business understanding to inform development of the Proof of Concept Model. The purpose of bringing together WPD, ElectraLink and IBM was to provide a shared understanding of end-users and their current processes related to LCT proliferation, enabling the development of outputs designed with users in mind. Additionally, the workshop captured business-led hypotheses around LCT proliferation and identified data that can be used to investigate them. Key WPD personnel were invited to the workshop, across a range of business areas, including innovation, network planning and policy.

The recommendation for innovation projects is that this workshop approach be adopted for future innovation projects to embed understanding of the project from the outset across business units, as well as to align the project with WPD business requirements.

Project start-up: The LCT Detection project used an Agile project management approach to technical delivery. To ensure expedient and smooth project mobilisation into delivery phase, particularly with data-driven projects, it is beneficial to invite the data scientists to the kick-off meeting as well as project managers, to facilitate cross-team and partnership understanding of e.g. the Sprint process, where for example a product or, in this case, proof of concept model, is developed through a number of iterative phases or ‘sprints’.

Daily touchpoints: The Agile approach to the technical delivery under the project meant that a series of daily touchpoints (10 – 15 minute telephone calls) were scheduled for the duration of the sprint process. The whole technical project team, led by the ElectraLink project manager and supported by ElectraLink’s technical lead, attended these calls with the IBM data science team and project manager. These daily calls ensured that any risk of data environment, technical platforms failure, or other issues were flagged as a risk at the earliest opportunity, with associated mitigations considered and reported back as required.

Business Values Report: This document was supplementary to the Sprint 2 and Sprint 3 Status Reports and was provided to WPD in order to ensure that the project continues to align with WPD’s business needs. Drafted and delivered by ElectraLink for the project, the Business Values Report reconfirmed the sound base for understanding the business requirements of the client in terms of what the PoC model needed to deliver.

The recommendation to WPD is that in future innovation projects, a Business Values Report is scheduled as a deliverable of projects where appropriate, to ensure continuing alignment with WPD requirements across the business.

Data analysis phase: Whilst we had multiple hypothesis around what we would see in the data, we organised a phase in the project to prove what we were trying to see before we commenced the larger piece of work. This ensured that the larger piece of work could be targeted to the datasets that provided the most value.

Operational process change: Given the relative paucity of EV, PV and other low carbon technology-related words in freeform text in the data sets, it would add value to the data and model output if engineers going out to properties recorded 'EV', 'PV, 'heat pumps' as a matter of course, in all cases, not just where issues are encountered.

It is recommended that this change is implemented under business as usual to support detection of LCTs from freeform text.

Dissemination activity: The project was launched at LCNI in November 2018, with an introductory four-page brochure produced for that event, to disseminate to DNO and energy industry stakeholders. A press release was issued to key energy media; this was used as a tool to engage stakeholders, with a news article being linked to from ElectraLink's website, on social media: https://www.electralink.co.uk/2018/11/low-carbon-technologies-lct-detection-project/. An on-going social media campaign, using Twitter and LinkedIn, has resulted in widespread interest in the project. A webinar was delivered in May 2019 to disseminate learning from the project. A summary brochure will be produced in June 2019 for dissemination to the whole DNO community and key industry stakeholders.

9 The Outcomes of the Project

9.1 Hypotheses

From the initial Design Thinking workshop – the project focused on two hypotheses.

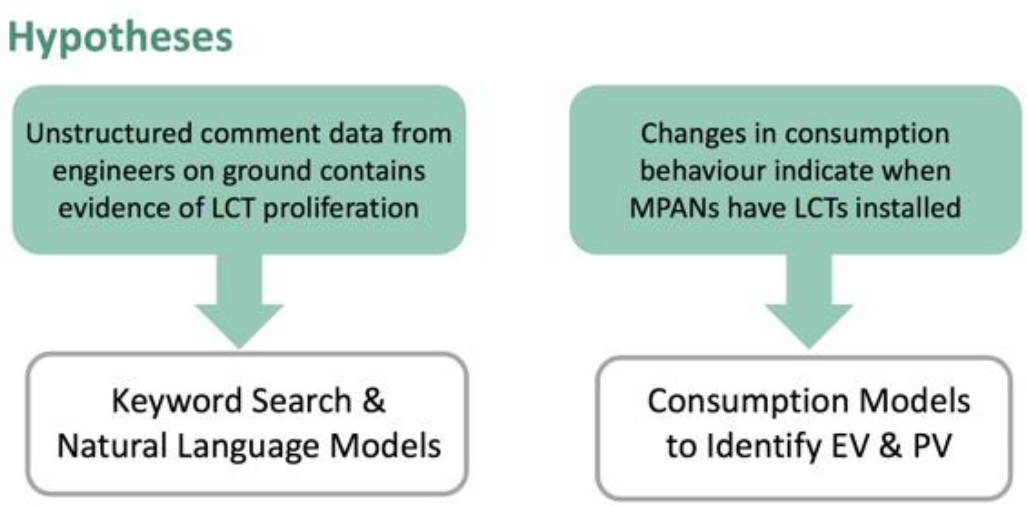


Figure 9-1 - Hypotheses

9.2 Unstructured Data Models – Initial Analysis

The DTS dataset includes comments made by engineers on call-outs written to help engineers on future visits

The comments are:

- Usually short sentences with a few words
- Often repeated

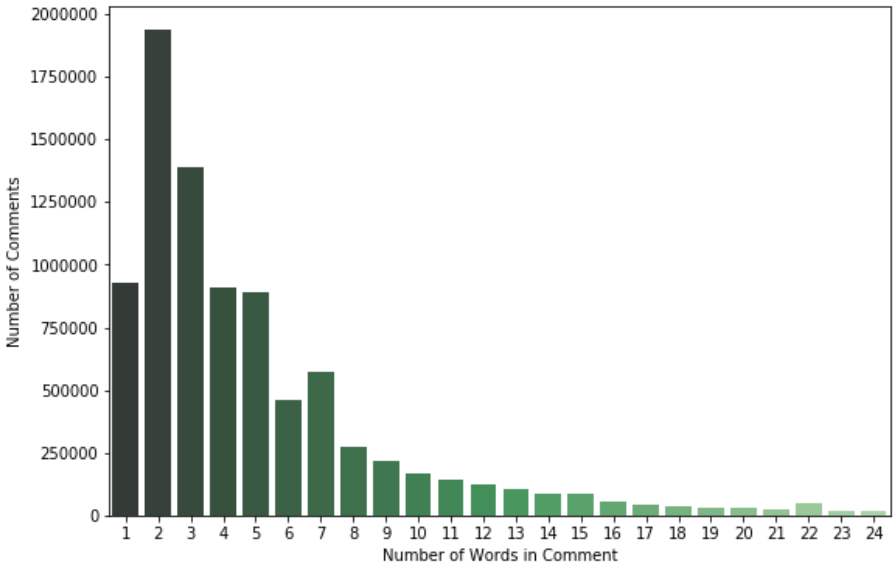


Chart 9-1 - Distribution of Word Count

Comment	# of times present	% of all comments
no access	387,518	4.38%
<blank>	338,927	3.83%
k	320,652	3.63%
standard service standard service levels	294,845	3.33%
white door	156,424	1.77%
nrq	142,644	1.61%
no answer	141,066	1.60%
unable to obtain a remote meter reading	117,175	1.33%
no more info	103,716	1.17%
brown door	49,481	0.56%

Table 9-1 - Most Frequent Comments

9.3 Unstructured Data Models – Natural Language

Natural Language Processing is concerned with the analysis of human language by an automated system. It attempts to understand meaning, sentiment, and other related aspects of the language present in the nuances we all use when communicating. Comments by engineers are often short phrases and not long full sentences, making them less suitable for a Natural Language Classifier. The Natural Language Classifier could not identify any new instances of LCT because the comments themselves did not contain enough evidence of LCTs' presence. Casting the human eye across the comments, it is difficult to recognise patterns to look for other than keywords related to LCT.

9.4 Unstructured Data Models – Keyword Search

Keyword searches related to LCT were undertaken to classify any MPANs with comments that contain these words. Evidence was found of meters running backwards caused by presence of photovoltaic cells – this would call for a field technician to visit the meter and hence there is a proliferation of comments associated with this issue.

Most positive matches for the keywords were related to PV. This is likely because of engineers being called out for unusual readings caused by the running backwards phenomenon.

Category	Keywords
EV	'charge point', 'electric vehicle', 'charging point', 'hybrid car', 'range extender', 'tesla', 'model x', 'model s', 'i3 ', 'i8 ', 'outlander', 'car plugged in'
PV	'voltaic', 'solar', 'pv ', ' pv'
Heat Pump	'heat pump', 'ground source'

Category	# of MPANs in tagged by keyword search	# of MPANs tagged by keyword search that are not in known dataset
EV	39	37
PV	3,042	1,924
Heat Pumps	4	4

9.5 Data

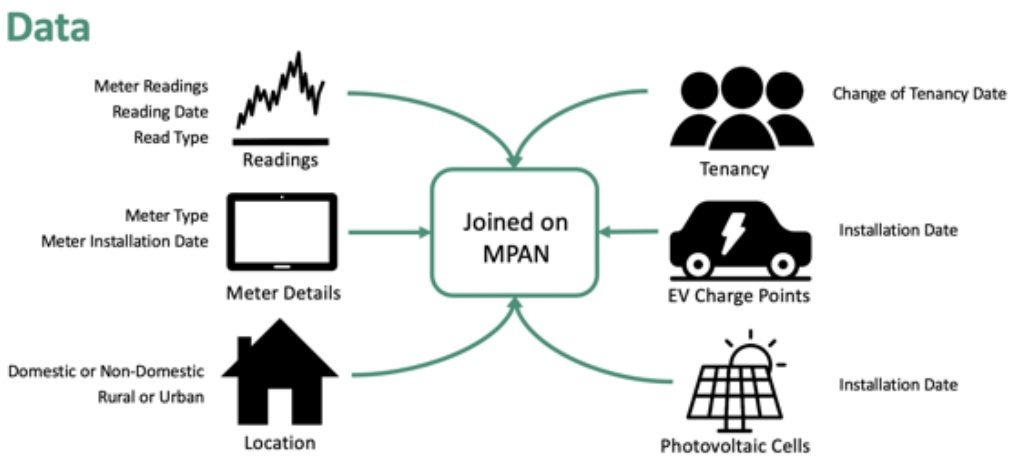


Figure 9-2 Data Sources

Multiple data sets were joined using the pseudonymised MPAN to create a master data set for Feature Engineering.

9.6 Data Processing

Further data cleansing was required prior to processing in order to remove "dirty" data items that could adversely affect the model. Examples of this:

- Anomalous spikes and drops in readings
- Duplicated readings
- Readings on same day

9.7 Feature Engineering

As detailed in 4.6.4 Initial Feature Engineering features were developed on the master data table to enable modelling. The key features include:

- Change since last reading
- Days between readings
- Average daily energy usage
- Change since same period last year
- Was LCT installed when the reading was taken?
- Has LCT been installed in last 12 months?
- Tenancy end dates

9.8 Analyses - Population-level analysis

Exploring the data through analysis and visualisations helps us to understand the data, enabling us to begin to build our approach to modelling. It also allows us to test some hypotheses:

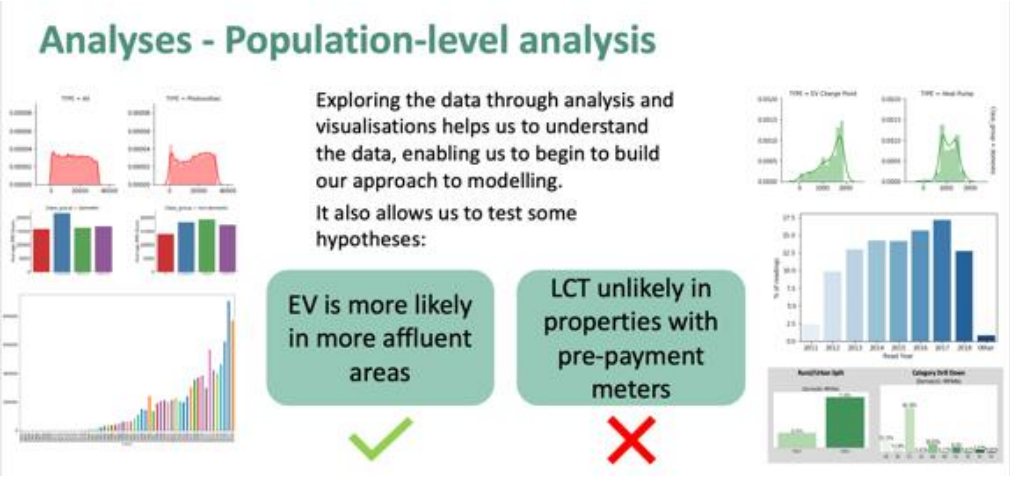


Figure 9-3 - Population Level Analysis

9.9 Analyses - Population-level consumption analysis

- When looking at multiple MPANs with LCT, the noise of individual MPANs' consumption is averaged out and we can see behavioural changes once LCT has been installed.
- In the majority of cases for PV, there is a reduction in energy consumption once PV has been installed.
- In the majority of cases for EV, there is an increase in energy consumption once an EV Charge Point has been installed.

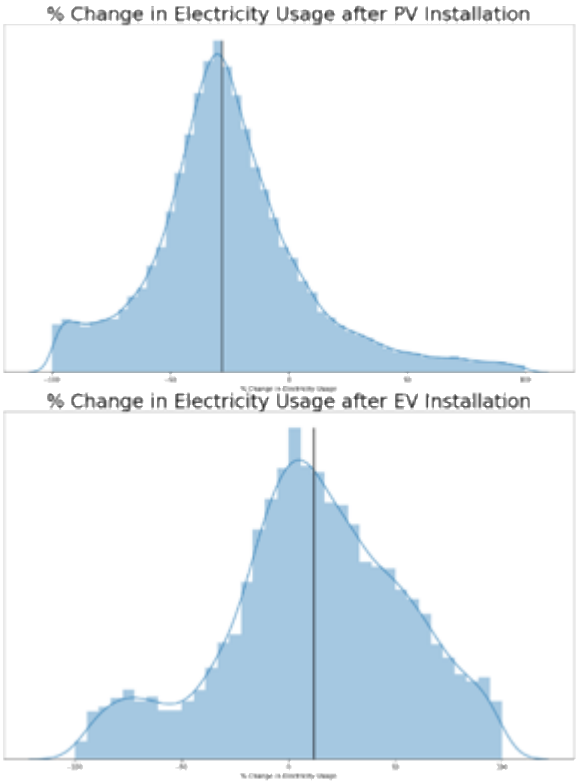


Chart 9-2 - Normal Distribution Skews - PV & EV

9.10 Consumption Model – Target Variable

Lack of a negative dataset (i.e. MPANs known not to have LCT) means:

- Demographic data cannot be used
- We cannot model on an MPAN level

Consumption is therefore the input for the model

Due to different meter reading frequencies average daily consumption was calculated, and due different levels of consumption, change in average daily consumption for a given MPAN (controlled for meter, register and tenant) was calculated. To account for seasonal effects, change in average daily consumption compared to the same period the previous year was calculated for each MPAN. Therefore, the target variable the model is trying to predict are the readings in the 12 months after LCT installation.

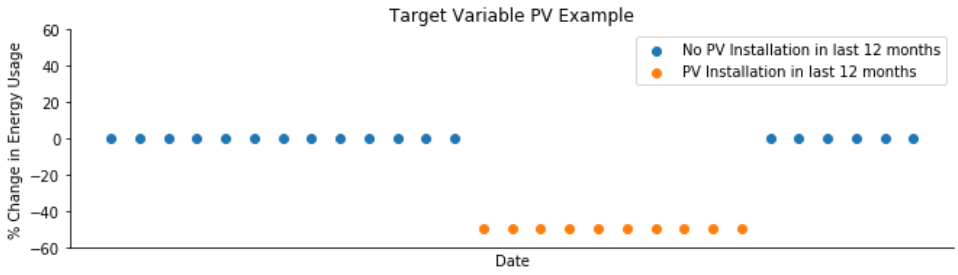


Chart 9-3 - Target PV Example

9.11 Initial Approach

The initial approach took all of the MPANs with the known LCT and built a model to predict LCT, however this was unsuccessful because:

1. Differences in the frequency of readings for different meter types, as well as difference between domestic and non-domestic. Therefore, the model was restricted to domestic smart meters.

MType_group	avg(DaysSinceLastReading)
Smart	37.023263419216 78
Others	35.188993267119 83
Prepayment	63.179315518696 45
Non-Smart	98.525267206961 77

Table 9-2 - Reading Intervals

2. A portion of the MPANs have a lot of noise from other unknown fluctuations in consumption

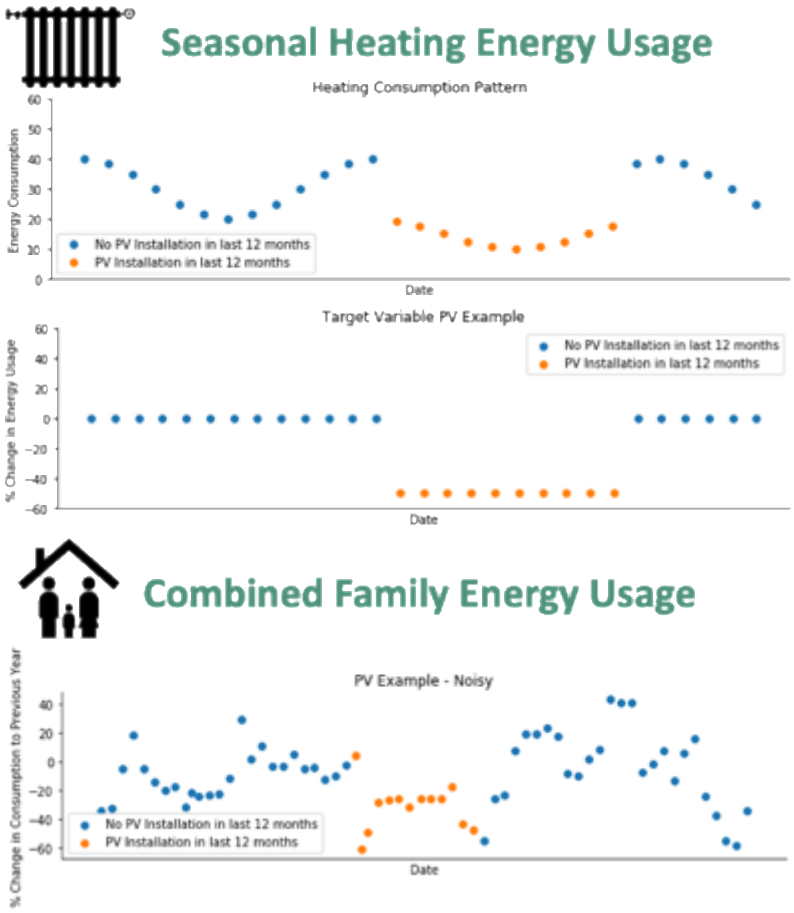


Figure 9-4 - Usage Patterns

9.12 Developing a New Approach

Based on the above findings, a new approach was developed:

1. MPANs in use were filtered to just domestic smart meters
2. Changed input to rolling average of consumption change compared to the same period last year, to smooth out the data
3. Selected MPANs with a strong and consistent signal from the known LCTs to build the model on.

$$rolling\ average = \sum_{i=0}^{i=90}(\Delta\ Consumption) \Big/ number\ of\ readings$$

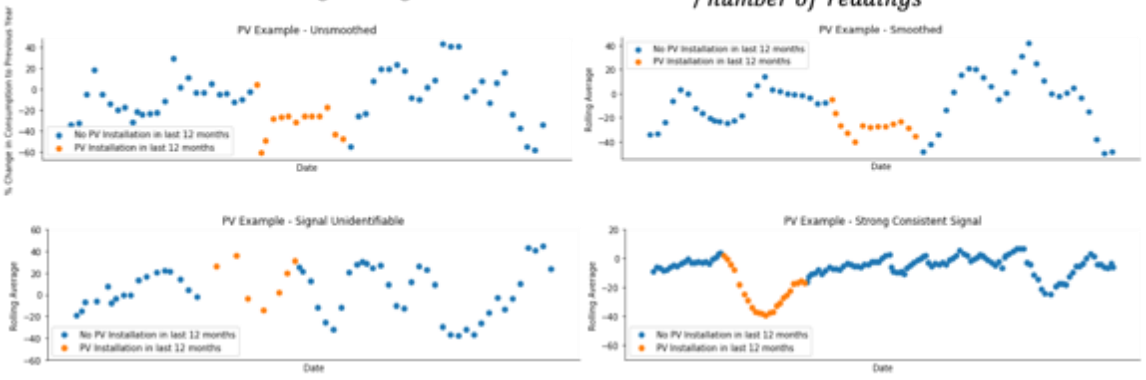


Chart 9-4 - Rolling Average Approach

9.13 Final Approach

The finalised approach to modelling with consumption data is described in the flow below:

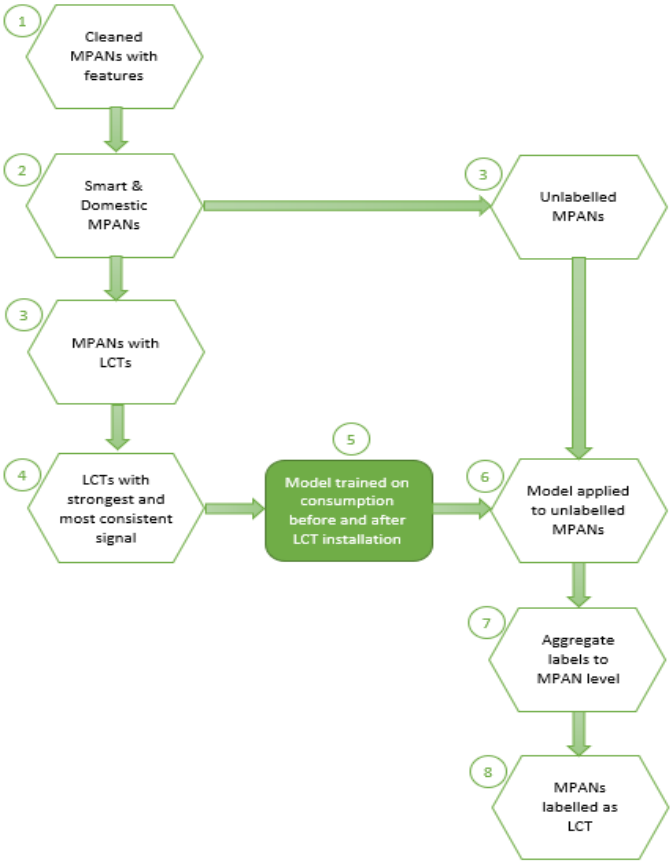


Figure 9-5 - Final Approach

1. Selecting MPANs for modelling. Domestic MPANs with smart meters known to have LCT are ranked by:
   a. Change in average daily consumption in the year before installing LCT compared to the year after
   b. Standard deviation of change in average daily consumption
2. Once these MPANs are ranked, a proportion of the MPANs with the largest change in consumption and the lowest standard deviation are used to train the model.
3. Selecting MPANs as predicted to have LCT. Two metrics were created to measure whether the aggregated predictions indicate whether an MPAN has LCT:
   a. Number of predicted periods of LCT / average number of readings per year
   b. Timespan between the first and last predicted period

MPANs meeting the following criteria are deemed to indicate LCT:

Metric	Thresholds
Predicted Readings / Average No. Readings per Year	0.9 < x < 1.1

Metric	Thresholds
Timespan between first and last predicted period (days)	250 < x < 400

Table 9-3 - Thresholds

9.14 Results

MPANs meeting the criteria in the final approach are deemed to show similar enough consumption behaviour to the MPANs with known LCT and a strong consistent signal to indicate they may also have LCT present.

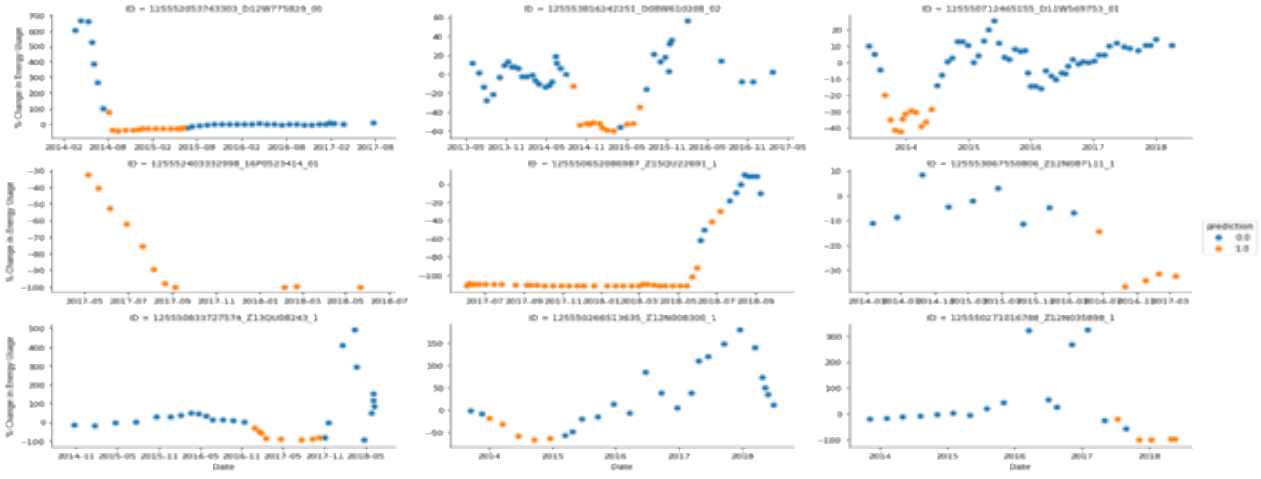
The number of MPANs identified for EV and PV are:

LCT Type	No. MPANs Predicted to have LCT
EV Charge Point	5,863
Photovoltaic	8,104

Table 9-4 - Predicted MPANs

9.15 Results - PV

Plots of a random selection of MPANs which the model predicts as having PV installed and which pass the thresholds:

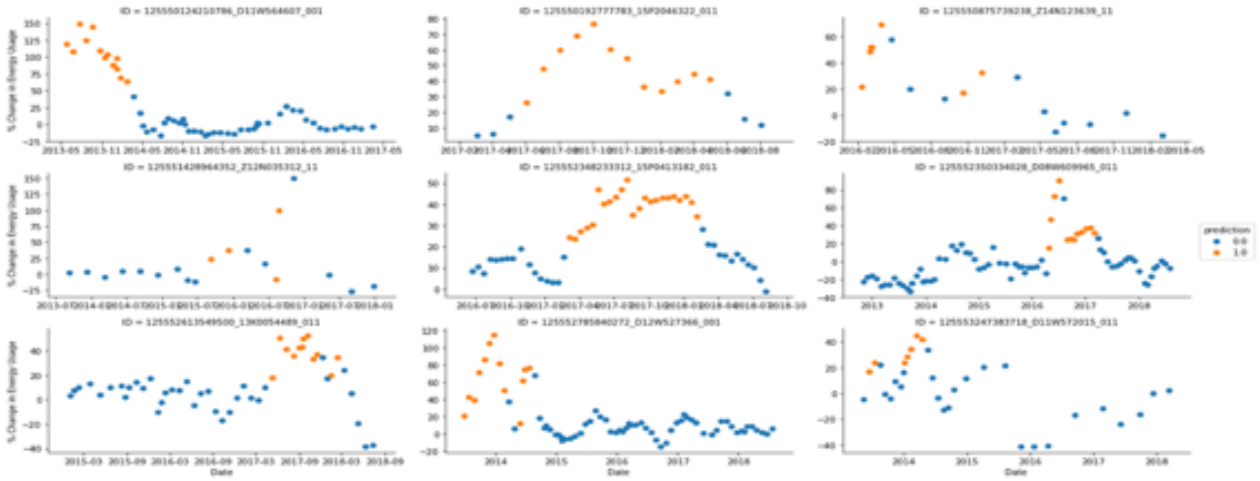


Plots of a random selection of MPANs which the model predicts as having PV installed and which pass the thresholds

Chart 9-5 - Results PV

9.16 Results - EV

Plots of a random selection of MPANs which the model predicts as having EV installed and which pass the thresholds:



Plots of a random selection of MPANs which the model predicts as having EV installed and which pass the thresholds

Chart 9-6 - Results EV

9.17 Dashboard

In addition to the modelling a dashboard was created to geographically display the locations of all the known LCT and LCT that the project had found. All code for the dashboard is available in Appendix A.

The dashboard is an interactive map based on a static data set derived from the outputs of the two models (unstructured data, consumption) as well as the known LCT from Crown.

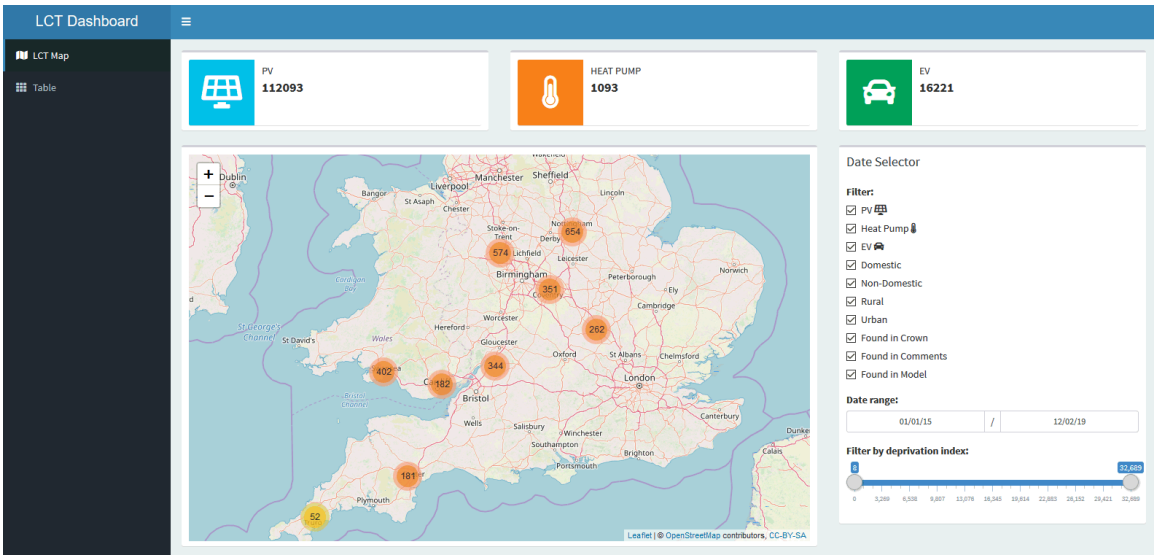


Figure 9-6 - R/ Shiny Dashboard

10 Data Access Details

The Python code required to implement the model is available as zipped Jupyter notepads and attached in Appendix A. However – a subset of the data relating to the LCTs identified in the project is attached as an Excel file.

The code (written in the language 'R') required to run the dashboard is attached in Appendix A.

The final result output of the model is contained in a single large CSV file which is available to WPD. This file is very large (~1 GB) and therefore not included as an attachment in this report.

The file contains the following fields:

COLUMN_NAME	MS_SQL_DATA_TYPE
id	bigint
pMPAN	bigint
Class_group	varchar
Meter	varchar
Register	varchar
MType	varchar
GEN_TYPE	varchar
GEN_Commissioning_Date	varchar
GEN_CAPACITY	float
LCT_TYPE	varchar
LCT_CAPACITY	float
LCT_Commissioning_Date	varchar
tenant_rank	bigint
IMD_SCORE	float
RURAL_URBAN_GROUP	varchar
latitude	float
longitude	float
MType_group	varchar
MSN	varchar
InstDate	varchar
EndDate	varchar
Found_in_CROWN	bigint
KEY_DATE	varchar
Found_in_Keywords	float
PV_Prediction_Date	varchar
Found_in_Model_pv	float
EV_Prediction_Date	varchar
Found_in_Model_EV	float
PV	bigint
EV	bigint
HeatPump	bigint
Found_in_Model	bigint

Table 10-1 - Final Data Output

11 Foreground IPR

The project has developed the following foreground IPR:

1. A keyword search programme to identify LCT presence from unstructured comments.
2. Data preparation and cleansing modules that take the DTS data set and produce features that enable modelling based on consumption data.
3. A binary random forest model that predicts LCT presence based on consumption data.

12 Planned Implementation

12.1 Phase 1

The consumption model developed can be reconstructed on an appropriate platform which provides:

- A Cloud Object Storage (HFS) component
- A Jupyter/ Python component
- An Apache SPARK component

The simplest way to implement the model would be to use the IBM Cloud Platform – however – with modification, other platforms which offer similar components can be used (e.g., Amazon AWS).

Within the Jupyter notebooks provided, there are clearly marked code cells which contain connection credentials to allow access to storage and to SPARK. These credentials will need to be updated to allow connection in a new environment.

The project team's recommendation is that the modelling phase is extended into a phase 2 / 3 which would focus on enhancing the model.

12.2 Phase 2

A second phase of the project would focus on the ability to harness unstructured data from ElectraLink, combine it with cognitive analytics, and use this to develop a cost-effective virtual monitoring capability, to create a digital twin, on WPD's network. There are two principal ways it will deliver value:

1. Enhancement of the proof of concept models under the LCT Detection project, together with other existing toolsets to solve visibility issues shared by WPD and other DNOs, predominantly focusing on lack of visibility of Electric Vehicles (EVs) and Low Carbon Technologies (LCTs)
2. By using this visibility and existing datasets, it will develop a mechanism to provide virtual monitoring capabilities in the absence of smart meters and access to smart data.

LCT Modelling: Understanding the network.

The first work package of this project will gather data on the network required for development of a Virtual Monitoring tool. All relevant existing datasets needed to develop the tool will be brought together, including WPD datasets (such as substation locations) and the DTS dataset (such as the Embedded Generation dataset). Other NIA project datasets will be analysed for model integration, e.g. Electric Nation, OpenLV, LV Connect and Manage.

An operational LCT detection model will be created to fill in the visibility gaps for EVs and LCTs on the DNO's network, using insights and learning from the LCT Detection project. The modelling will use meter flow messaging data to advance WPD's understanding of LCT locations. DNOs have mechanisms in place through which they should be informed about Distributed Energy Resources (DERs) that are connected to the network (e.g. G83, EV Charge Point Installation Notification), but the information provided through these means is incomplete (as highlighted during the LCT Detection proof of concept). Our hypothesis is that we can significantly enhance the LCT modelling to predict the presence of LCTs by analysing the following data:

- Consumption data
- Historic weather data
- Known generation
- Demographic data
- Freeform text comments made by engineers
- Notified installations (e.g. G83)

With the rich history of data ElectraLink has captured over many years, the benefits to forming a more accurate picture of the prevalence of LCTs are:

- Improved DNO power flow models to predict network loading (e.g. reverse power flows) and power quality issues.
- Model the potential for demand management / curtailment options.

Trials: Understanding the power flows and creation of half-hourly load profiles

Current lack of access to half-hourly (HH) data about household power flows on the WPD network inhibits understanding of where EVs and LCTs are connected at LV level. This results in the need to install physical monitoring at substations. The VM DATA project will investigate the feasibility of creating HH load profiles for EV / LCT households that can be fed into a Virtual Monitoring tool. The output will be analysed against actual substation monitoring data, to determine if profiles can be used as a proxy for direct access to smart meter data for the purposes of network monitoring. To create these load profiles, WPD will install HH meters at identified households (c. 200) to develop consumer-type load profiles (non-LCT Urban; LCT suburban, rural etc.) to understand if this can be used to improve consumer profiles and minimise the need for HH data. Substation GridKey and Voltage monitoring will be installed to validate the virtual monitoring.

Business as Usual: Virtual Monitoring Tool

The final stage of the project is to take the output and learning from the first two work packages – understanding what is on the network and understanding the impact this will have on power flows – and create a 'digital twin' for WPD's network. This 'digital twin' can

then be used by WPD to understand its network constraints and identify any requirement for more granular monitoring.

What would the second phase project deliver?

The 'VM Data' project will deliver a virtual monitoring (VM) capability on WPD's network. This will reduce need for physical monitoring, avoiding the costs associated with physical monitoring and transformer replacement; supporting better phase balancing and demonstrating RIIO-ED2 cost savings and transition to DSO.

The second phase project will give WPD greater insight into its network and processes by integrating a richer mix of structured and unstructured data and processing it at speed. This virtual monitoring capability will be a key requirement as DNOs transition to DSO; there will be many more variables to optimise and data to incorporate.

The move to production is illustrated in the table below.

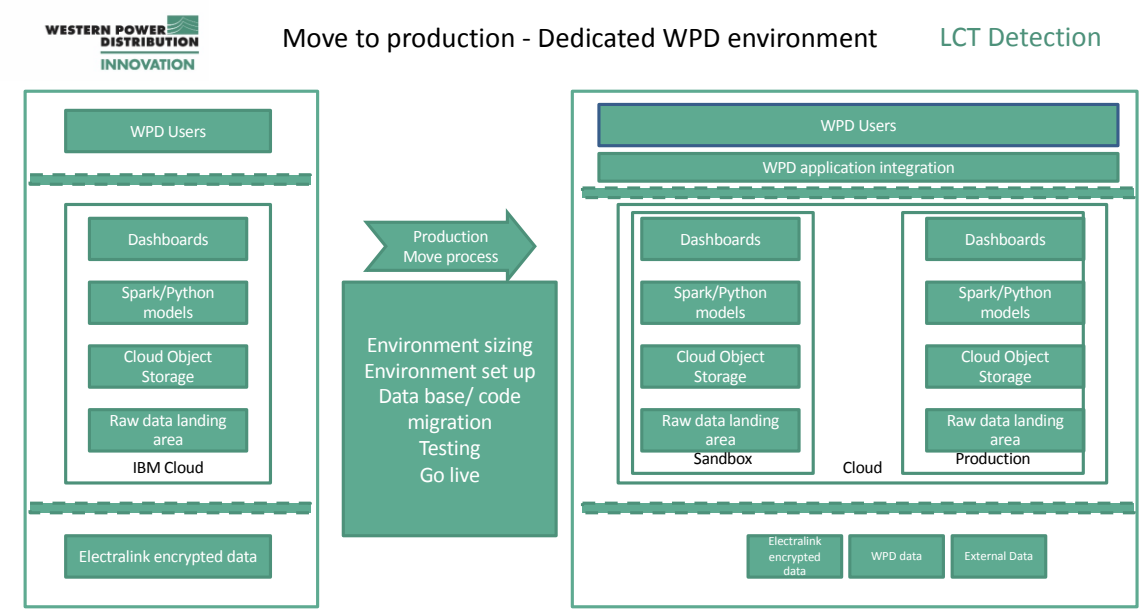


Figure 10-1 - Move to Production

ElectraLink and IBM believe that the second phase project will deliver a BAU-adaptable technology platform that will drive benefits for WPD, its network and customers in the future through digitalisation of processes to enable facilitation of EV and LCT uptake.

13 Contact

Further details on replicating the project can be made available from the following points of contact:

Future Networks Team
Western Power Distribution,
Pegasus Business Park,
Herald Way,
Castle Donington,
Derbyshire
DE74 2TU
Email: wpdinnovation@westernpower.co.uk

14 Glossary

Abbreviation	Term
LCT	Low Carbon Technology
EV	Electric Vehicle
PV	Photo-Voltaic, referring to solar installations.
BAU	Business As Usual
DNO	Distribution Network Operator
DSO	Distribution System Operator
MPAN	Meter Point Administration Number
LV	Low Voltage
DTS	Data Transfer Service
DER	Distributed Energy Resources
HH	Half Hourly
VM	Virtual Monitoring
PoC	Proof of Concept

Table 14-1 Glossary

15 Appendix A

15.1 Python Code

LCT Detection_Python Code.zip

15.2 R/ Shiny Dashboard Code

LCT Detection project_R Shiny Dashl

15.3 Identified LCT Extract

LCT Data Extract.xlsx