faculty

SUBAK

Fellowship Report

Tori Pereira

7 June - 19 July 2021

Tech mentor: Ashwin Chopra

Commercial mentor: Greg Jackson

Combining Socioeconomic and Energy Datasets to Forecast Electric Vehicle Uptake

Client Overview

Subak is a non-profit accelerator for data-driven organisations addressing climate change. As an accelerator, Subak will provide commercial and technical support to their member start-ups. A key aspect of Subak's mission is to coordinate a data cooperative in which participating organisations share their climate data to support stronger data insights, identify climate risks and opportunities, and to measure climate impact.

The fellowship project involved using data provided through one of Subak's accelerated start-up's NewAutomotive. NewAutomotive is an independent transport research organisation who aim to support the transition to Electric Vehicles (EVs) in the UK through marketing, technology and policy.

Project Scope

The Faculty fellowship project sought to connect disparate datasets relevant to EV uptake in the UK. A key aim was to understand if we could develop a forecast of EV uptake in the UK driven by socioeconomic and energy factors such as income.

The key objectives were:

1. Create a pipeline taking in publicly available datasets and outputting a dataset containing the features and target variable (EV count) on a granular discretisation.
2. Forecast EV growth in local regions of the UK.
3. Understand what the best predictors (socioeconomic, charger distribution, relative price) for EV growth are.
4. What is the impact of EV usage on CO2 intensity?

The Data

Established granularities subsetting the UK:

Middle Super Output Areas (MSOAs)
Lower Super Output Areas (LSOAs)
Local Authorities (LAs)
Postcodes (PCDs)

Publicly available data:

1. Household income - MSOA
2. House price - LSOA
3. Rural-urban classification (RUC) - LSOA
4. Index of multiple deprivation (IMD) -LSOA

5. Electricity consumption - LSOA
6. Photovoltaic (PV) solar panel count - PCD
7. Public chargers - LA
8. Government grants for private chargers (interpreted as a count of private chargers) - LA

NewAutomotive vehicle data:

9. MOT tests for all vehicles (including EVs)
10. MOT test centre locations - PCD

Approach and Results

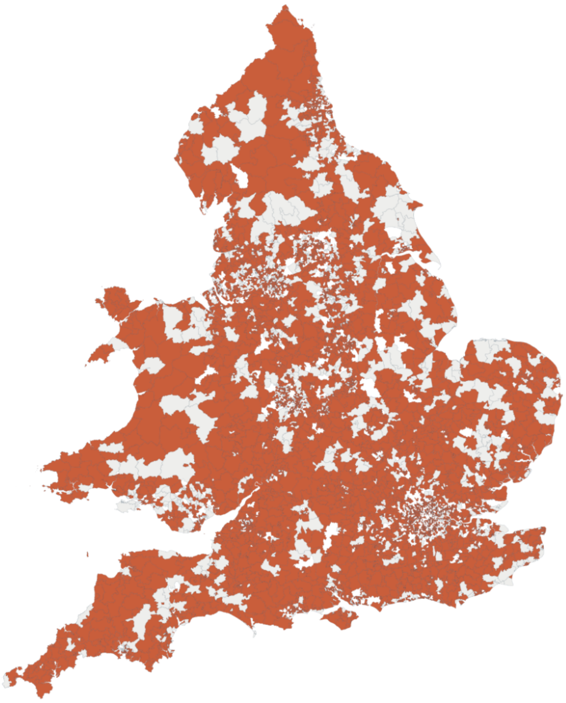
Data cleaning and processing

The first three weeks of the project involved accessing and processing the data. We chose to consider the granularity of MSOAs in England and Wales. England and Wales, as we could find publicly available data on all the feature variables for these two countries. As we were using the MOT test centre data for our target variable, we considered only the MSOAs in England and Wales that had test centres located within them. This ensured that MSOAs with zero EV count were true zeros, and this left us with 5740 MSOAs.

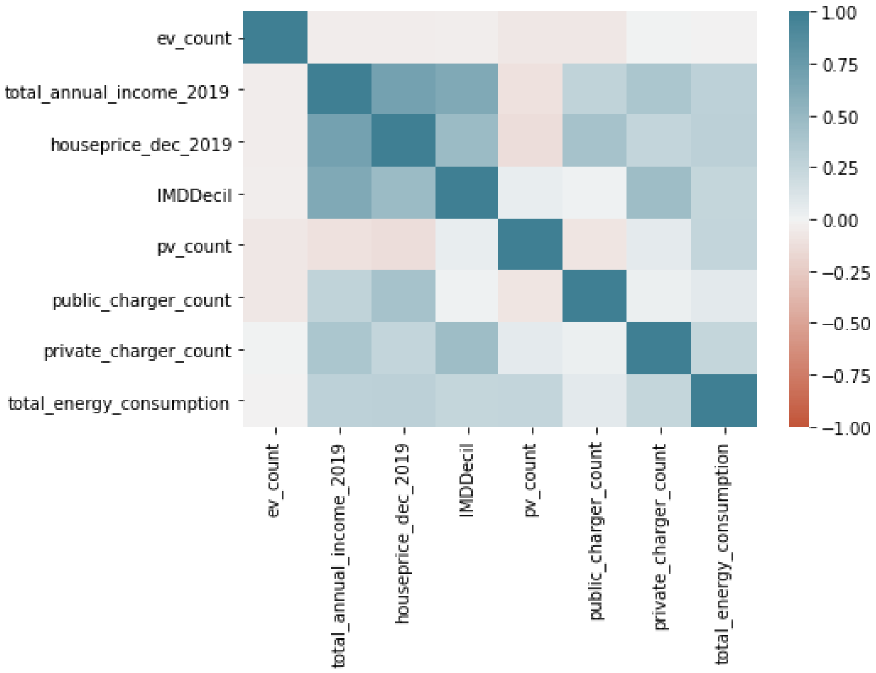
As advised by NewAutomotive we took all the socioeconomic features (income, house price, RUC, IMD, electricity consumption) to be constant in time, as this avoided having to include a forecast for these variables. The time-dependent variables were the PV solar panels, and public and private charger counts. The target variable provided through NewAutomotive was the count of EVs in each MSOA, as given by the MOT test centre data.

Exploratory data analysis

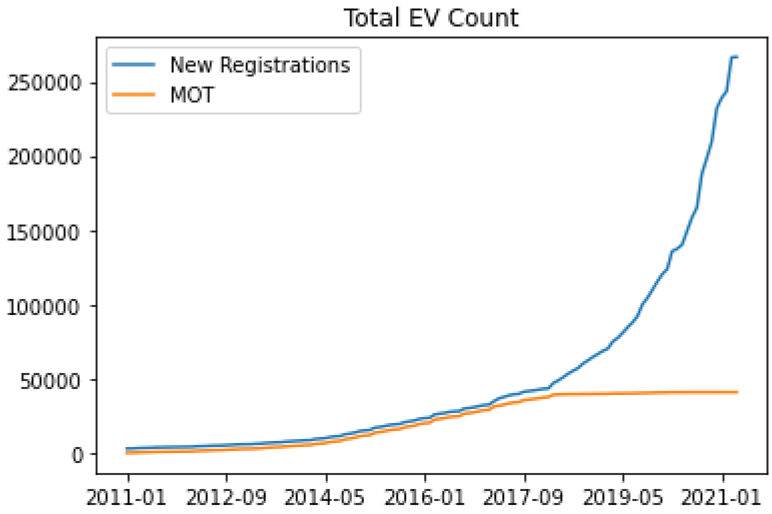
Before developing any models, we completed some basic exploratory data analysis. There was a heavy class imbalance between the number of MSOAs 4922 that have EVs present (coloured in orange below), and 818 (coloured in light grey below) that do not.



We found very little correlation between the features and target variable:



To assess the accuracy of the MOT data for a proxy of EV location, we compared the total sum of the EVs across both countries as given by the MOT data to that given by new registrations (see figure below). This showed that the MOT data significantly undercounts for EVS in the last three years. This is likely due to the fact that new cars only need to have their first MOT when they are three years old.

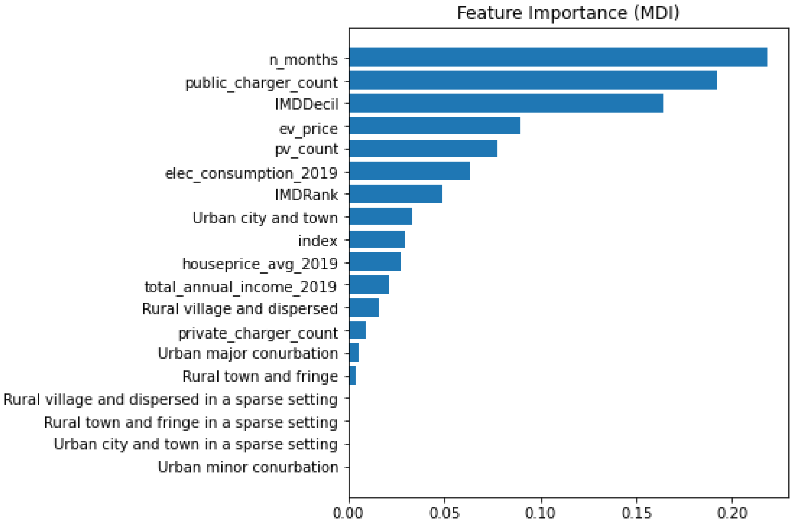


Modelling

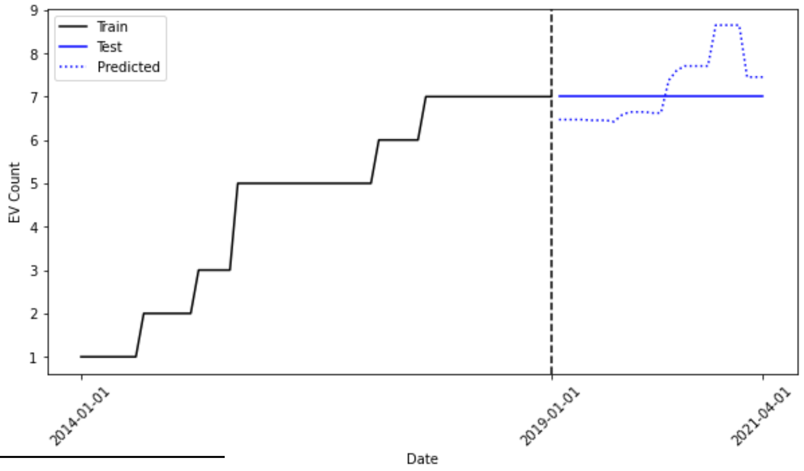
1. Forecast

The exploratory data analysis of the EV count showed that the MOT data we had did not provide an accurate value for our target variable. Nevertheless, we developed Random Forest and XGBoost models to forecast the EV count using both the steady and time-dependent features. The XGBoost model was consistently more accurate than the Random Forest so we chose to refine the XGBoost model.

We included an additional *n_months* variable counting months from the earliest time stamp. The resulting model had very low accuracy (~0 for both training and testing). One useful result is the feature importance as shown below. We found that public charger count was the most important feature. This is consistent with recent studies exploring the factors affecting EV uptake [1,2].



To improve the forecast, we include an additional variable *ev_avg_2014* which is the average number of EVs in the first year 2014. The resulting forecast had a significantly improved accuracy (0.99 R2 training and 0.91 R2 testing), and for this version of the model the most important feature was the new variable *ev_avg_2014*. An example forecast for a single MSOA region is shown below.



[1] Christidis, Panayotis, and Caralampo Focas. "Factors affecting the uptake of hybrid and electric vehicles in the European Union." *Energies* 12.18 (2019): 3414.

[2] Fevang, Elisabeth, et al. "Who goes electric? The anatomy of electric car ownership in Norway." *Transportation Research Part D: Transport and Environment* 92 (2021): 102727.

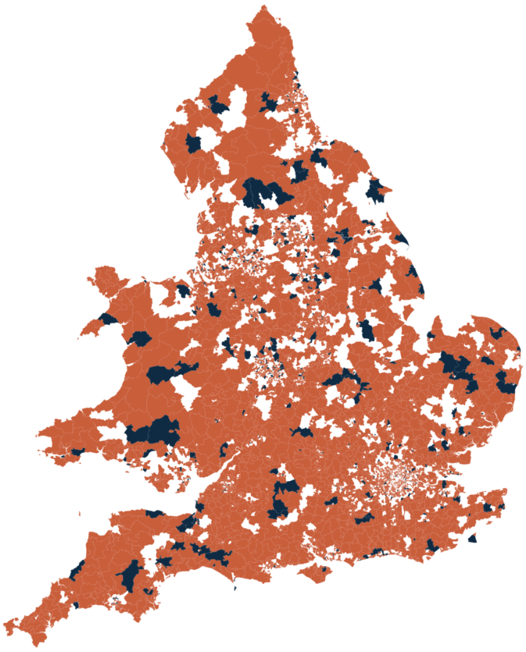
The example demonstrates a problem with using tree-based models to forecast which is that these models do not enforce nonnegative growth. Thus, future work for improved models could include autoregression and recursive time-stepping forecasts.

2. Classification

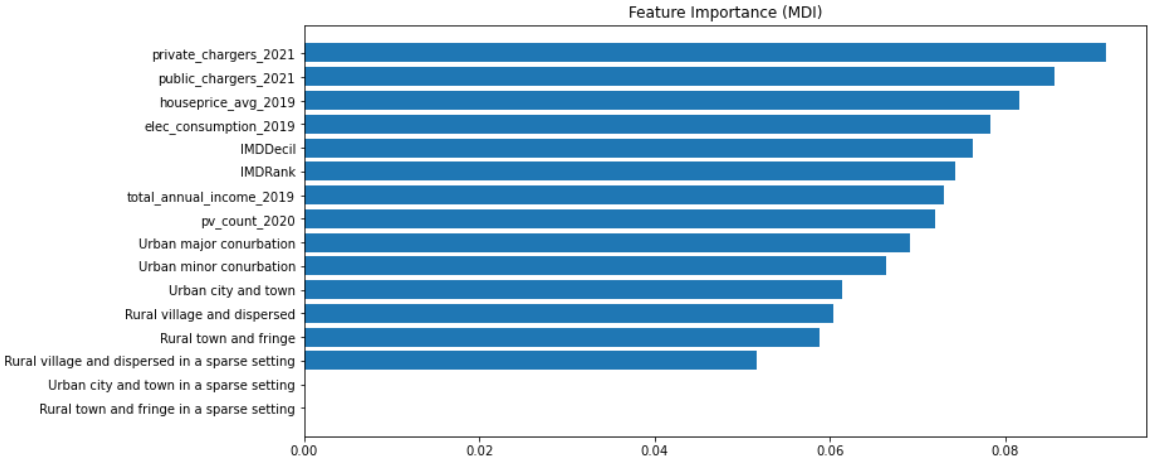
Motivated by the weakness in forecasting capabilities, we refocused the project to ask the question: are there regions that currently don't have EVs (according to the MOT data) but are predicted to do so from the features? We call such regions *untapped*. To address this, we built a classifier to predict the presence of EVs. We trained on 70% of the MSOAs and tested on the remaining data. The results are summarised in the table below.

Test set results	Predicted: No	Predicted: Yes
Actual: No	(TN) 7	(FP) 385
Actual: Yes	(FN) 5	(TP) 1326

We can interpret the false positives (coloured blue below) predicted by the classifier as untapped regions and compare the features of the false positives to the true positives (coloured in orange below). This is because the classifier has learnt from the training data that these regions, according to their socioeconomic and energy feature values, should have EVs present.

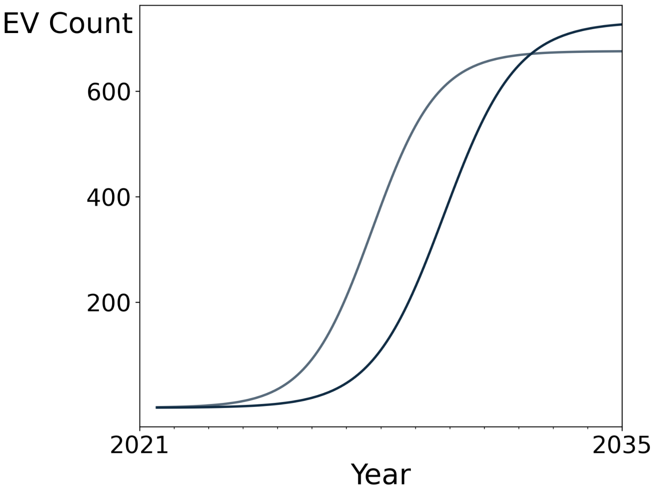


As with the forecast, we can look at the importance of the features for the classification. Here the most important features were found to be the private and public charger counts and the average house price; which aligns with intuition.



3. Back to the Forecast (Logistic growth model)

We used the classification results to propose a simple logistic growth model for the forecast. We created a dataset detailing the number of cars in each MSOA using the MOT data for all cars in 2020. We used this to give a 'capacity' for each MSOA, and assumed logistic growth between the number of cars there are in the MSOA today (as given by the MOT data) to the 'capacity' which we assume must happen by 2035. This left two parameters for fitting the curve (the midpoint and growth rate). The growth rate was assumed to be proportional to the probability of EVs as predicted by the classifier, and the midpoint was chosen to be earlier for regions which had higher probability of EV presence as given by the classifier. We show the resulting modelled forecasts for two MSOA regions below. The curve that grows faster had a higher probability of EV present than the slower growing curve.



Challenges Encountered

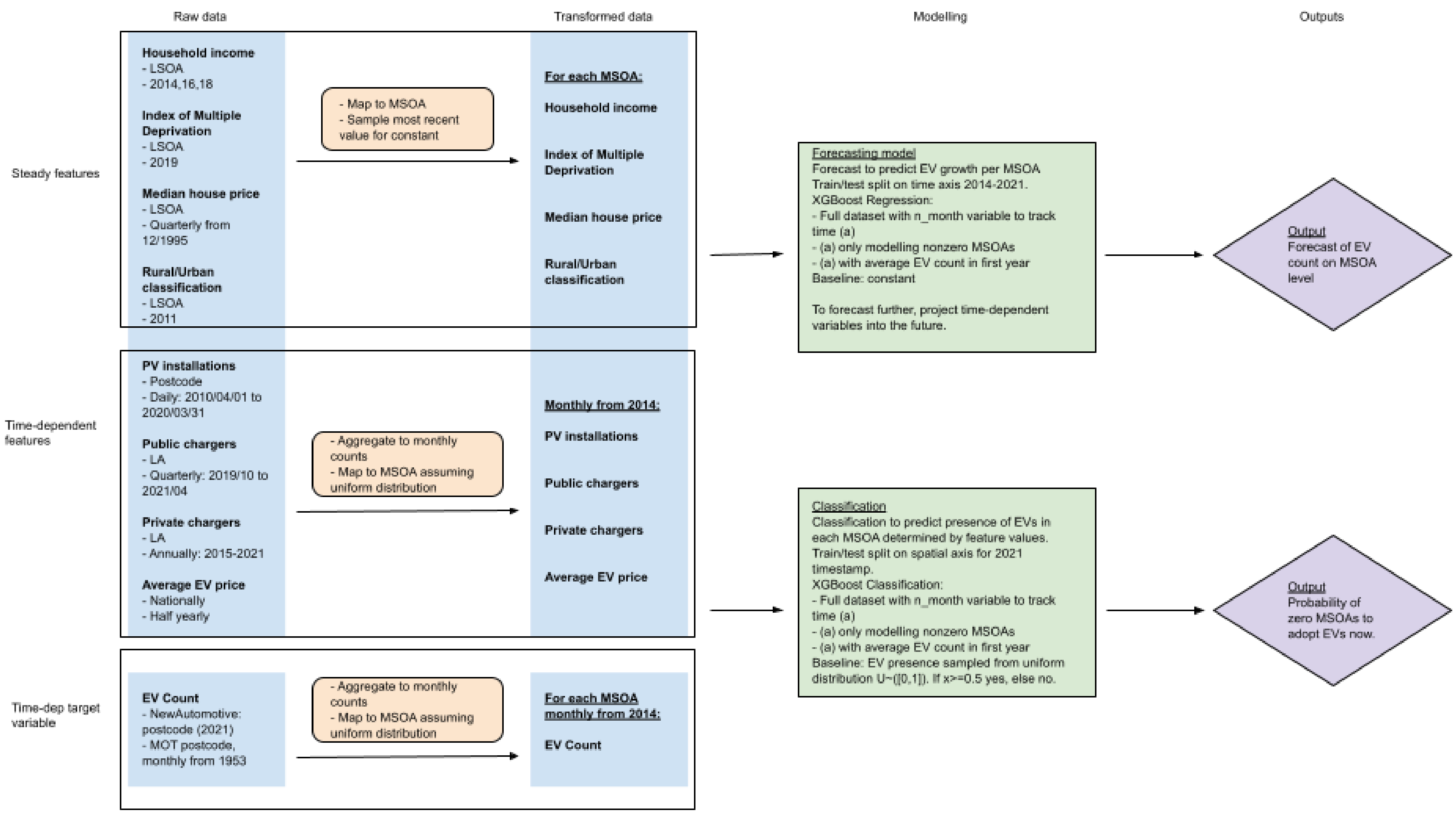
- The biggest challenge was that we did not have accurate data for the target variable of EV count. The available MOT data is not accurate for the past three years due to delay in getting the first MOT.
- Likely due to the lack of accuracy with the target variable data, we found EV count to be uncorrelated with the predictive features. This meant simple multivariate linear regression would have been totally ineffective at forecasting.

Impact

This project provides Subak with a pipeline to access, clean, and connect datasets for a number of socioeconomic and energy variables (e.g. income, public and private charger availability) measured across the UK at local regions producing an open, combined dataset that will be published in the data cooperative. Moreover, two machine learning models have been developed. The first is a classification model which provides predictions as to whether a particular region has yet adopted EVs. This model gives insight and predictions for the regions which do not yet have EVs but are on the brink of adoption. The second model is a forecasting model for EV uptake in each local region.

The data is publicly available on figshare at:
https://figshare.com/articles/dataset/MSOA_evcount/14995020

The code is publicly available of Subak's github:
https://github.com/ClimateSubak/EV-forecasting



Raw data

Transformed data

Modelling

Outputs

Steady features

Household income
- LSOA
- 2014,16,18

Index of Multiple Deprivation
- LSOA
- 2019

Median house price
- LSOA
- Quarterly from 12/1995

Rural/Urban classification
- LSOA
- 2011

- Map to MSOA
- Sample most recent value for constant

For each MSOA:

Household income

Index of Multiple Deprivation

Median house price

Rural/Urban classification

Time-dependent features

PV installations
- Postcode
- Daily: 2010/04/01 to 2020/03/31

Public chargers
- LA
- Quarterly: 2019/10 to 2021/04

Private chargers
- LA
- Annually: 2015-2021

Average EV price
- Nationally
- Half yearly

- Aggregate to monthly counts
- Map to MSOA assuming uniform distribution

Monthly from 2014:

PV installations

Public chargers

Private chargers

Average EV price

Time-dep target variable

EV Count
- NewAutomotive: postcode (2021)
- MOT postcode, monthly from 1953

- Aggregate to monthly counts
- Map to MSOA assuming uniform distribution

For each MSOA monthly from 2014:

EV Count

Forecasting model
Forecast to predict EV growth per MSOA
Train/test split on time axis 2014-2021.
XGBoost Regression:
- Full dataset with n_month variable to track time (a)
- (a) only modelling nonzero MSOAs
- (a) with average EV count in first year
Baseline: constant

To forecast further, project time-dependent variables into the future.

Classification
Classification to predict presence of EVs in each MSOA determined by feature values.
Train/test split on spatial axis for 2021 timestamp.
XGBoost Classification:
- Full dataset with n_month variable to track time (a)
- (a) only modelling nonzero MSOAs
- (a) with average EV count in first year
Baseline: EV presence sampled from uniform distribution U~([0,1]). If x>=0.5 yes, else no.

Output
Forecast of EV count on MSOA level

Output
Probability of zero MSOAs to adopt EVs now.