EPC and Heating Type Predictions

Below is the technical documentation for the project segment that focused on predicting both the EPC ratings and the heating type of homes that are not currently in the EPC database. This document is broken down into the following sections:

1. Data
2. Machine Learning Model Selection
3. Similarity Quantification (SQ) Model
4. Combining Models
5. Performance After Including Proxies
6. Recommendations for Future Work

This document will not include a detailed discussion of the data preprocessing steps. For that discussion, please see the *Preparing EPC Data* document.

During this process, we made several assumptions. We have highlighted those assumptions in red.

Data

We used the following data sources for predicting the EPC ratings and heating sources of the homes not in the EPC database:

- EPC Database
- Fuel Poverty Data
- Energy Consumption Data
- Ordinance Survey Maps

See the *Getting Proxies* and *Preparing EPC Data* technical documentation for more details on the preprocessing and data selection process.

Machine Learning Model Selection

We performed a rigorous model selection process to determine which models could best predict the EPC ratings and the heating sources. We tested with Naïve Bayes, Support-Vector-Machine, Random Forest (RF), AdaBoost, and XGBoost. We also used a model based on home similarity, discussed in section 3.

We tuned the model's hyperparameters using cross-validation before comparing the models against each other. You can find the tuning results in APPENDIX A: Model Output Metadata. We compared the models using several metrics: accuracy, root-mean-squared error (RMSE), area under the precision-recall and ROC curves, and the macro-F1 scores. You can find details on these metrics here.

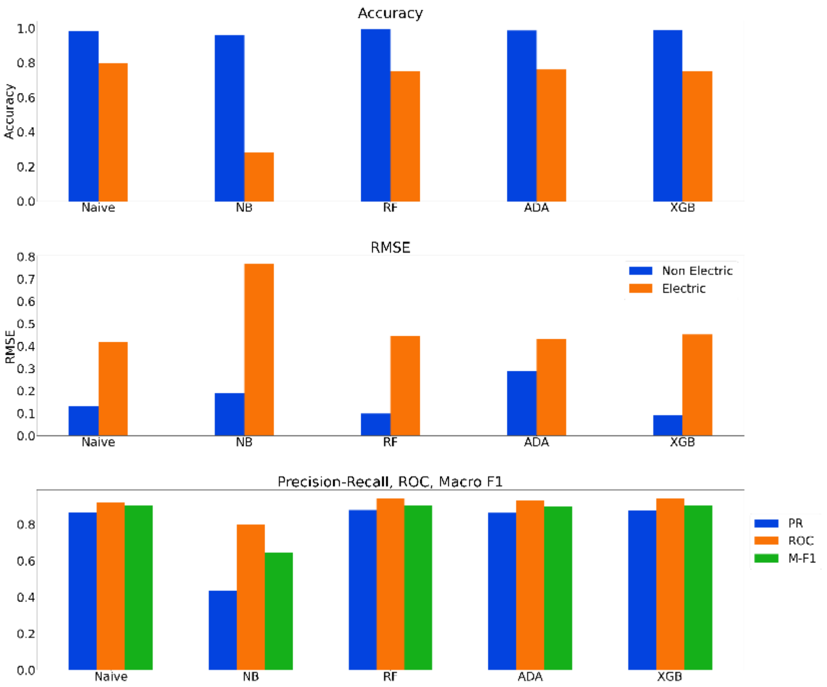

*Figure 1 Results for the heating type predictions. The top row shows the accuracy scores (perfect score of 1) for the five models. The second row is the RMSE (perfect score 0), the final row is the area under the precision-recall and ROC curves, and the M*acro *F1 scores.*

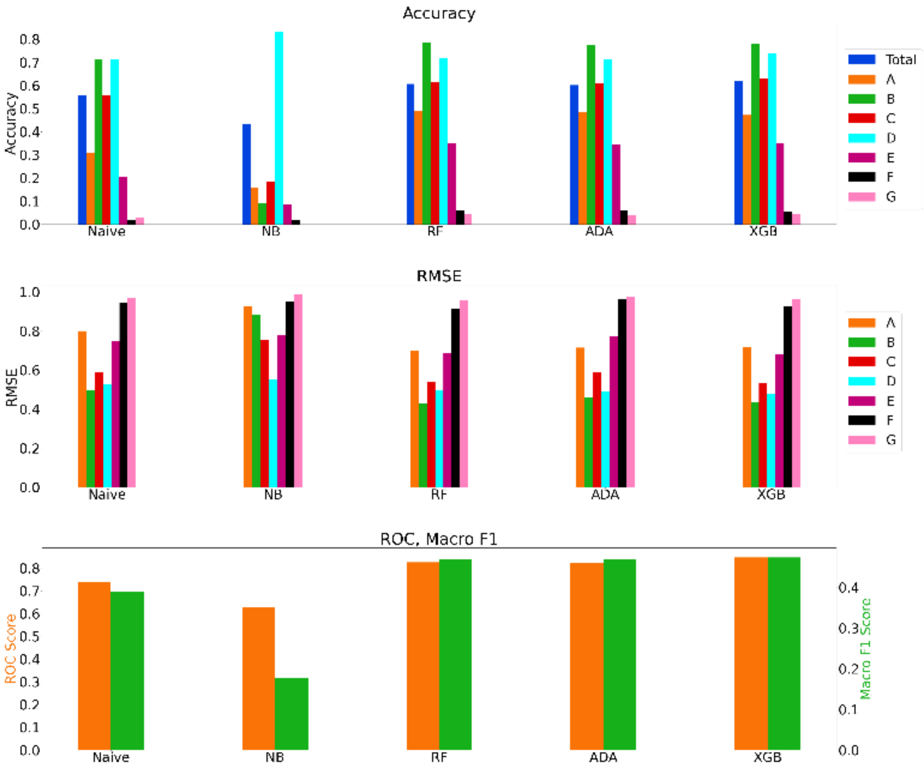

Figure 2 Results for the EPC predictions. The top row shows the accuracy scores (perfect score of 1) for the five models. The second row is the RMSE (perfect score 0), and the final row is the area under the ROC curve and the Macro F1 scores.

The RF, ADABoost, and XGBoost models performed similarly for both the EPC and the heating type predictions. Since all three models follow a similar methodology, we chose the RF model because it has a shorter training time.

The models' output is a probability representing the share of decision trees in the RF that make the same prediction. The model outputs one probability for each of the classes: the seven EPC ratings (A-G) or the two heating types (electric or non-electric). We take the class with the highest probability as the predicted value. We also use the probability as the level of confidence in the predictions to provide a confidence level that the true value of the predicted EPC is within one rating of the predicted rating.

Our partners are interested in locating optimal areas for intervention and retrofitting. The idea is that retrofitting large groups of homes together in the same location is a more cost-effective method of improving homes' energy efficiency, thus reducing their carbon emissions. In our model performance analysis, we examined the accuracy as a function of how many homes in the postcode have the same rating. Figure 3 shows the results of this analysis.

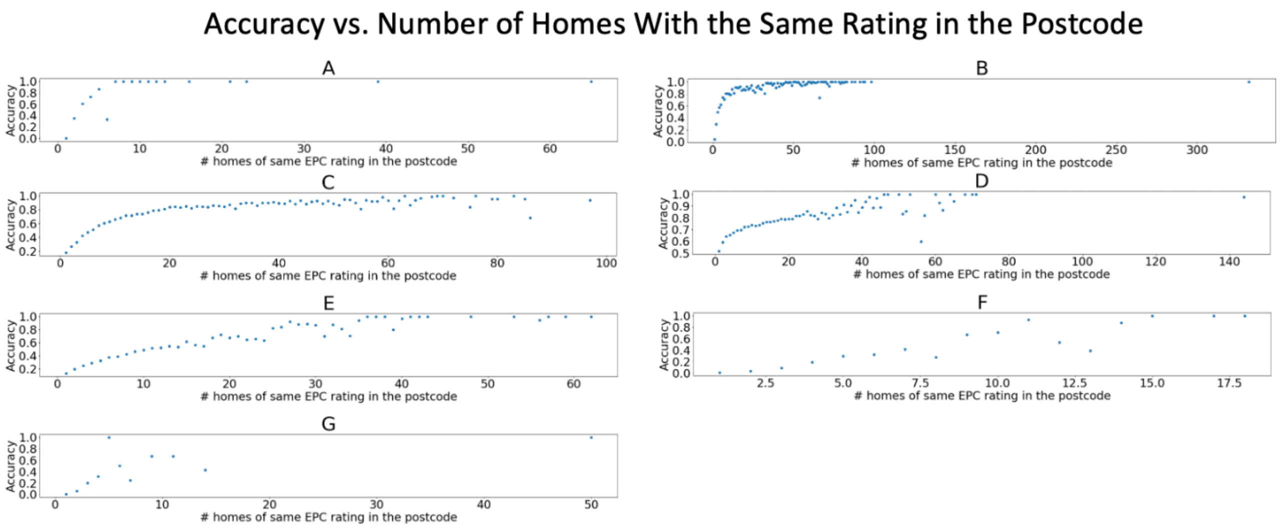

Figure 3 Model accuracy as a function of how many homes in the same postcode have the same EPC rating.

As can be seen, the model accuracy increases as the number of homes in the postcode with the same rating increases. The models are skilled at predicting groups of houses with the same ratings, which may be one of the reasons the models are less accurate for the F and G bands. Homes in the F and G bands appear to be more spread across geographic areas and rarely grouped. Figure 3 shows the results for this part of our analysis. As can be seen, the accuracy tends to be low where there are few homes of the same type in the postcode and rises as the number increases.

Similarity Quantification (SQ) Model

We assumed that homes built at the same time by the same builder in the same area would have the same EPC rating. Thus, we developed a method to look at homes in a defined area and compare the shapes of houses not in the EPC database with homes in the EPC database. The ratings from the homes with EPC ratings would be mapped to homes of similar shape that are missing ratings. The figure below shows the shape of homes in the EPC database. The colour of the shapes indicates the EPC rating of the homes.

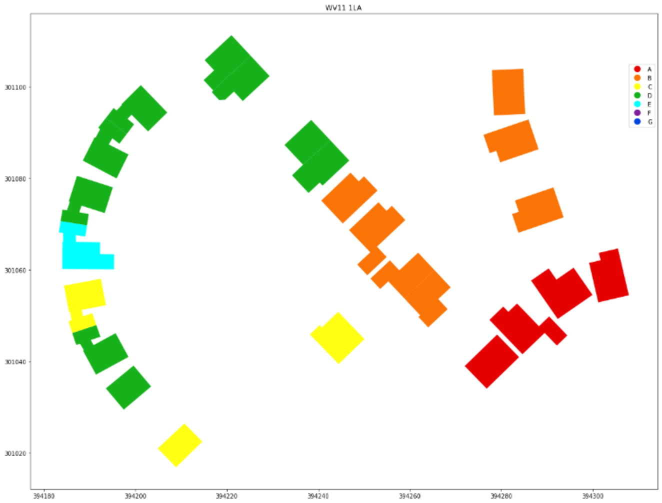

Figure 4 Homes with an EPC rating in a postcode. The colour represents the EPC rating, created using the .shp files for the homes in this postcode from the Ordinance Survey Maps.

The OS maps contain a shape file for each home in the target area. We attempted to compare these shapes directly, but we could not develop an effective method for performing this on a large enough scale to analyse all the homes in the West Midlands. Instead, we used each house's geometry to calculate each house's floor footprint area using geopandas. We calculated the area to various levels of precision, and two decimal places produced the best overall accuracy. The homes were compared within the same postcode, LSOA code, and MSOA code, with the postcode level producing the best overall accuracy. There were many homes with the same floor footprint area with multiple matches in the same postcode, with the matches not always being of the same rating. When this occurred, we took the most common rating of all the matches.

We derived a confidence level for this Similarity Model using a series of rules that depend on (1) how many homes in the postcode match the home area and (2) the differences in the ratings of the matching homes. If a home has multiple floor footprint area matches in the same postcode, and the most common rating of the matching homes makes up more than 66% of the ratings, the confidence of that prediction will be 0.8. If only one home matches the floor footprint area, the confidence rating drops to 0.5. If the most common rating of the matching homes lies between 34% and 66%, the confidence rating will be 0.5 and below 34% produces a confidence rating of 0.3.

The testing phase of this model indicated the overall accuracy of the SQ model is 0.62, which outperforms the RF model's accuracy of 0.55. Breaking down the accuracy into the different EPC bands reveals an interesting phenomenon. The SQ model has a very high accuracy for the A band, and the accuracy steadily declines as we move down the EPC ratings, ending at 0.04 at the G band. The similarity hypothesis being more accurate for homes with higher EPC ratings could be due to several factors. Homes built more recently, and thus having higher ratings, may not have had enough time for differences in owner behaviour to appear in different EPC ratings. Future work would benefit from focusing on the differences in these results and the underlying causes. The full results for the SQ model can be seen in Figure 5 in the section on combining models below.

Combining Models

While the performance of the SQ model is quite high for many of the different EPC bands, it is only able to provide predictions for those homes which have a match in the examined area. This limits the homes for which it can make predictions. Thus, the model must be combined with the RF model, producing a combined output.

Combining the models was done in the following steps:

1. The SQ model was used to get the ratings for ***4%*** of homes with at least one match, and the RF model predicted the EPC ratings for ***all*** homes in the testing dataset.
2. Populate the combined energy efficiency prediction if:
    a. SQ prediction = RF prediction: Keep prediction and RF confidence level
    b. SQ prediction != RF prediction AND RF confidence <0.5: Keep SQ prediction and the SQ confidence
    c. SQ prediction != RF prediction AND RF confidence >0.5: Keep RF prediction and RF confidence

The results of this method can be seen in Figure 5. The final model results are significantly lower for many of the EPC bands than for the SQ model alone. However, the combined model can slightly improve the RF results, which still ensures results for the whole testing dataset. The metadata for the final output files can be found in APPENDIX A: Model Output Metadata.

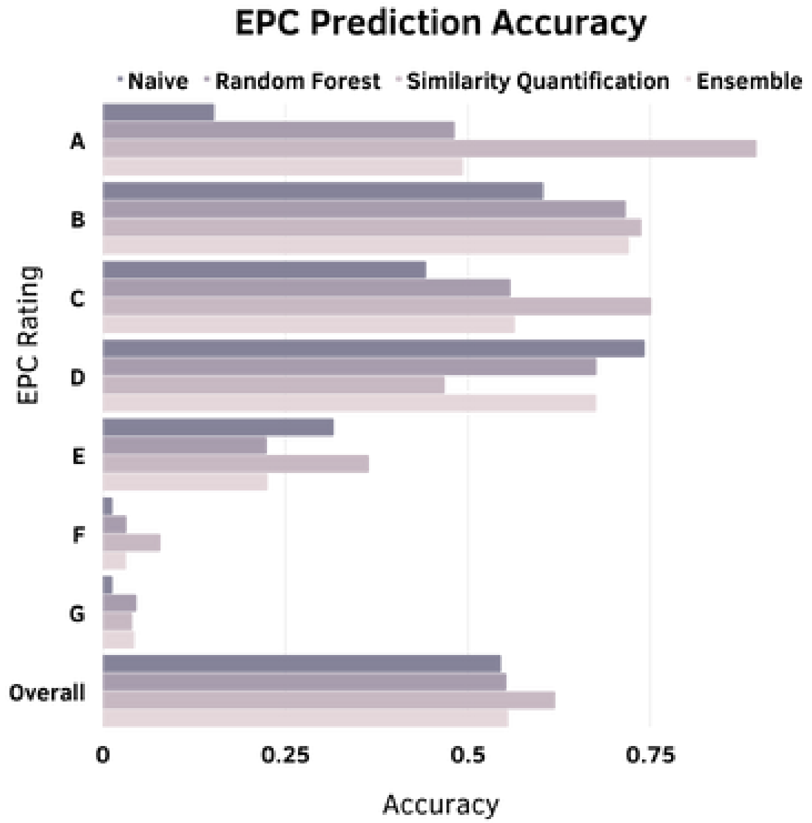

Figure 5 Results of the Naïve, SQ, RF, and the Combined models. Represents the accuracy for the different ratings and the overall accuracy once they've been combined.

Performance After Including Proxies

We initially trained and tested the model using data from the EPC database including values we expected to be able to get proxies for including the total floor area and the floor height.

- **Total floor area**: Derived from the building geometry included in the OSMaps for approximately 700,000 homes in the West Midlands
- **Floor height:** Derived from the relative maximum height (the distance between the ground and the highest point of the building's roof) from the OSMaps data

The other data used as input to the models as described in section 1, is either an identifier for the particular home, or is area level data (i.e., the mean consumption is the average energy consumption for all the homes in the LSOA area).

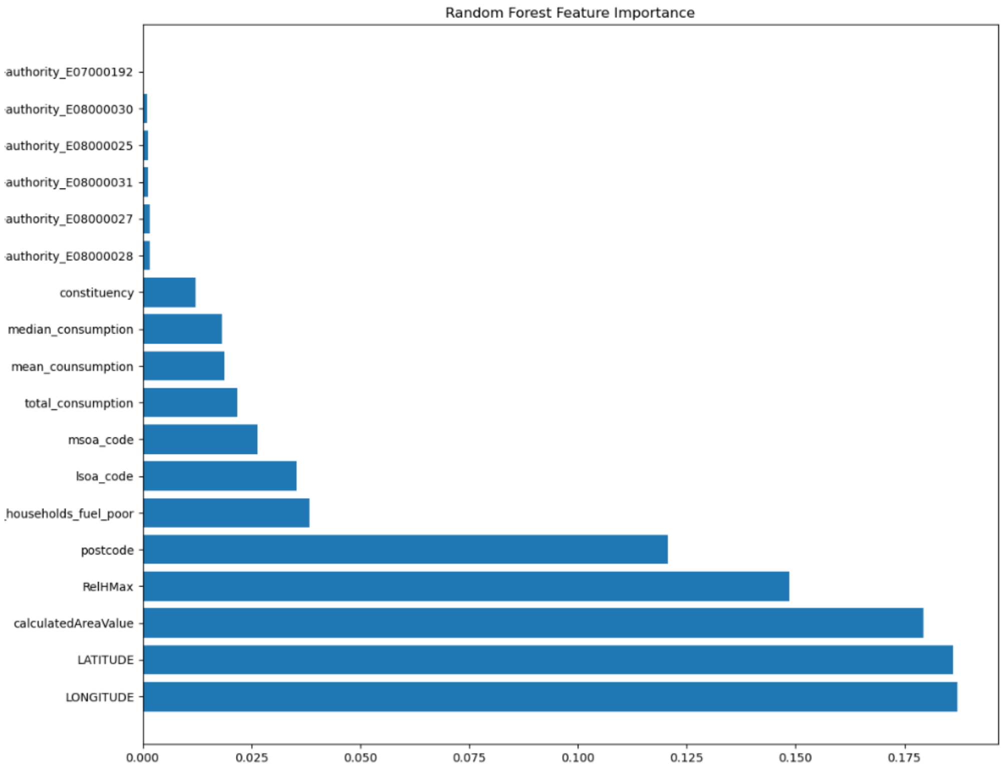

Figure 6 Feature importance for the final RF model. Latitude and Longitude come out as the highest, followed closely by the 'calculatedAreavalue'.

The overall accuracy of the RF model dropped by about 0.05 after replacing the proxies with real data. These results put the final RF model only about 0.02 points above the Naïve model. The Naïve model simply assigned the most common rating in a postcode to the homes that did not have a rating.

This serves to reinforce how influential the area-level data is to the results of the models. The feature importance in Figure 6 below shows how influential these area values are in making the EPC predictions. Increasing the accuracy of the models far beyond the baseline will require more house-level data. The total floor area and building height are influential factors but not sufficient on their own to provide perfect models.

When we initially tested all the data from the EPC dataset, the Energy Consumption data was by far the most predictive. However, this data is lacking on an individual house level for homes outside the EPC dataset. Future work would benefit from expanding the set of data unique to each home which may help differentiate the homes within a given area.

Recommendations for Future Work

We hope this project will serve as a baseline on which others can take and improve the models and predictions, better enabling the necessary work to reduce carbon emissions from building usage. We hope that some of these suggestions for future work based on our experiences with this project give a head start to those that continue this work.

More Advanced Models

Due to computing power and memory limitations, we had to limit the size and complexity of our machine learning models. The final RF model only contained 300 estimators, which limited its predictive power. Future work should consider using machines with more computational power and RAM to train more complex models that may improve the results.

We found that the accuracy improved as we increased the number of estimators in the RF model. This improvement was seen most significantly between different orders of magnitude, I.e., going from 100 to 1000 showed significant improvement in accuracy, while going from 100 to 500 showed minimal improvement. We could not test even higher numbers of estimators, but it is feasible that going up to 10,000 or more estimators would show a significant jump in accuracy. The number of estimators can be changed by tweaking the *n_estimators* parameter in the *multiclass_randomforest.py* file.

Google Streetview Images

We made an early attempt to analyse the similarity between homes using images collected from Google Streetview. The idea is the first and most ambitious iteration of the eventual SQ model discussed in the Similarity Quantification (SQ) Model section.

However, we were limited by funding since we had to pay a significant amount to pull all the images required for our region of interest. Thus, we did not explore this option further. It is possible that using these images would allow for an expansion of our SQ work. While quite accurate, it was only able to provide ratings for a small percentage of homes in the West Midlands. This work discusses the use of Google Streetview in EPC predictions.

Energy Consumption Data

The energy consumption data was the most predictive factor in the initial models we created in the exploratory phase. This is to be expected given that less efficient homes would consume more energy while more efficient homes would consume less.

Energy consumption is not a parameter available for individual homes outside the EPC database. However, because of this parameter's predictive power, we attempted to replicate the value. We used an iterative imputing method on the EPC data to see if it could be accurately replicated. However, the results indicated this method was not viable. Exploring other methods of replicating energy consumption could lead to some of the most significant improvements in these models.

Main Heat Description Pipeline

While cleaning the data, we converted the main heat description from its descriptive format into a binary value indicating electric or non-electric heating. We did this manually, with the unique values examined and sorted, largely relying on the methodology from Sonia William's *Using machine learning to predict energy efficiency*.

These descriptions vary in their level of detail, spelling, and method of description since they are input manually by inspectors. Moreover, with the EPC data being updated every six months, future entries likely may not match the descriptions we have manually sorted. A more thorough pipeline for sorting these values would be essential for reproducing this work in the future.

Land Surface Temperature

One method we considered for analysing the energy efficiency of homes was examining infrared satellite images, which can infer the temperature of the object in the image. The idea was that if we could obtain infrared images of homes on days that were either particularly hot or cold, we could compare the temperature of a house's roof with the surrounding land. If the temperature was significantly different, we could assume the home was less insulated and thus less efficient. However, we could not obtain infrared images of sufficient resolution to perform this analysis. This level of analysis would most likely be very predictive if these could be obtained.

APPENDIX A: Model Output Metadata

Column Name	Column	Type	Source
uprn	UPRN	Int	AddressBase Premium
postcode	Postcode	String	AddressBase Premium
LATITUDE	Latitude	Float	openuprn
LONGITUDE	Longitude	Float	openuprn
lsoa_code	LSOA code	String	ONS UPRN Directory
msoa_code	MSOA code	String	ONS UPRN Directory
prop_households_fuel_poor	Proportion of households in fuel poverty	Float	Sub-regional fuel poverty data
total_consumption	Total energy consumption within the LSOA code	Float	LSOA domestic electricity consumption data
mean_consumption	Mean energy consumption within the LSOA code	Float	LSOA domestic electricity consumption data
median_consumption	Median energy consumption within the LSOA code	Float	LSOA domestic electricity consumption data
constituency	Parliamentary constituency in which the building is located	Encoded Int	EPC database
calculatedAreaValue	Area per floor (m^2)	Float	OSMap Topology
local-authority (encoded)	Local authority area in which the building is located.	Encoded columns	EPC database
RelHMax	RelHMax = AbsHMax – AbsHMin (total building height)	Float	OSMap Building Height Attributes

current-energy-rating	EPC rating, either predicted or from the EPC database	String	EPC database or predicted value (see predicted column)
current-energy-efficiency	Continuous version of the EPC rating	Float	EPC database or predicted value (see predicted column)
SQ_current-energy-rating	Rating the house achieved from the SQ model. If nan, no rating was given by this model	String	Predicted (models/similarity_quantification_model.py)
SQ_confidence	Confidence of the SQ prediction. If nan, no rating was given by this model	Float	Predicted (models/similarity_quantification_model.py)
RF_current-energy-rating	Rating the house achieved from the RF model. If nan, the home has an existing EPC rating	String	Predicted (models/multiclass_randomforest.py)
RF_confidence	Confidence of the RF prediction. If nan, the home has an existing EPC rating	Float	Predicted (models/multiclass_randomforest.py)
confidence	Confidence of the final prediction	Float	Predicted by either the SQ or RF model, chosen in the models/combining_SQ_and_RF_results.py file)
confidence_within_one_rating	Confidence that the prediction is within one rating of the final prediction. Only produced by the RF model. If nan, final prediction was not made by the RF model or has an existing EPC rating.	Float	Predicted (models/multiclass_randomforest.py)
mainheat-description	Heating type. 0 if non-electric heating, 1 if electric heating.	Binary	From the EPC database or predicted.

additional_load	Calculated yearly kWh of power that will be put onto the network by the home if it were to switch from a non-electric heating type to a heat pump.	Float	Calculated (models/combining_results_for_output.py)
additional_peak_load	Calculated peak kWh of power that will be put onto the network by the home if it were to switch from a non-electric heating type to a heat pump.	Float	Calculated (models/combining_results_for_output.py)
predicted	Binary flag indicating whether the current-energy-rating, current-energy-efficiency, and mainheat-description were in the EPC database or were predicted. If 0 the values were in the EPC database, if 1 they were predicted.	Binary	Assigned (models/multiclass_randomforest.py)