Energy effectiveness of a mortgage portfolio

Jakub Vondráček, Daniel Till

2.1.2023

Overview

1. Data analysis
   - Data exploration
   - Augmenting the dataset
   - Datasets
2. Models
   - Used models
   - Model evaluation
3. Results
   - Imbalanced data
   - Predefined distribution
   - Feature Importance
   - Balanced data
4. Appendix

Data exploration

Value	Count	Frequency
D	8650444	43.8%
C	5789245	29.3%
E	3602636	18.3%
F	919699	4.7%
B	476051	2.4%
G	281359	1.4%
A	11011	0.1%

Table 1: CURRENT_ENERGY_RATING – frequencies in the cleaned dataset.

$$A < B < C < D < E < F < G$$

Client has the following data available:

- `BUILDING_AGE_CLASS`,
- `FLOOR_AREA`,
- `NUMBER_HABITABLE_ROOMS`,
- `PROPERTY_TYPE`,
- `POSTCODE (ADDRESS)`.

Augmenting the dataset

- Left out the variables: POSTCODE.
- Included the variables:
  - A_PROP_POSTCODE, . . . , G_PROP_POSTCODE,
  - LOCAL_AUTHORITY_LABEL,
  - BUILT_FORM,
  - POSTCODE_PROPORTIONS_ARE_RELIABLE_IND,
  - IS_EPC_LABEL_BEFORE_2008_INCL,
  - POSTCODE_COUNT.

Datasets

- Small dataset - client given data + augmentation.
- Big dataset - all available features.

Splitting the dataset

- Training set - 25 % of the observations.
- Validation set - 25 % of the observations.
- Test set - 50 % of the observations.

Adjusting class weights

- Imbalanced dataset.
- Predefined distribution (predefined by the client).
- Balanced dataset.

Used models

- Ordinal regression
- Support vector machine
- Gradient boosted decision trees

Model evaluation

Used metrics

- Accuracy
  - Proportion of correctly classified examples
- Ranked Probability Score
  - Metric that takes into account ordinality
- Confusion matrix
  - Precision, Recall, F1 Score

Imbalanced data

	Light Gradient Boosting Machine Big	Baseline Model Small	Light Gradient Boosting Machine Small	Ordinal Regression Small	Support Vector Machine Small
A	0.311	0.413	**0.358**	0.369	0.405
B	0.093	0.247	**0.127**	0.132	0.154
C	0.040	0.088	0.056	**0.054**	0.077
D	0.029	**0.027**	0.032	0.035	0.041
E	0.060	0.113	0.080	0.097	**0.061**
F	0.081	0.259	0.170	0.203	**0.162**
G	0.071	0.421	**0.293**	0.343	0.295
Weighted Avg	0.042	0.082	**0.060**	0.066	0.067

Table 2: Ranked Probability Score for models with imbalanced dataset.

Light Gradient Boosting Machine Big	Baseline Model Small	Light Gradient Boosting Machine Small	Ordinal Regresion Small	Support Vector Machine Small
0.657	0.438	**0.565**	0.533	0.519

Table 3: Accuracy for models with imbalanced dataset.

Imbalanced data

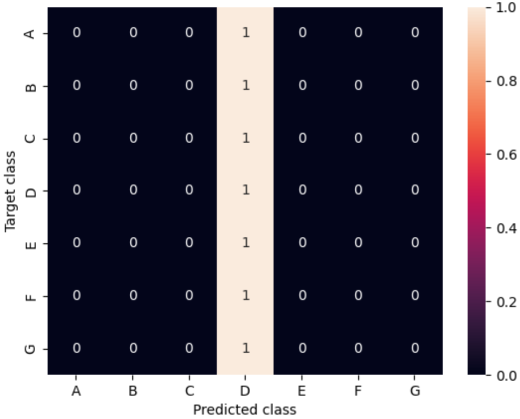

Figure 1: Confusion Matrix for Baseline Small Model with imbalanced dataset.

Imbalanced data

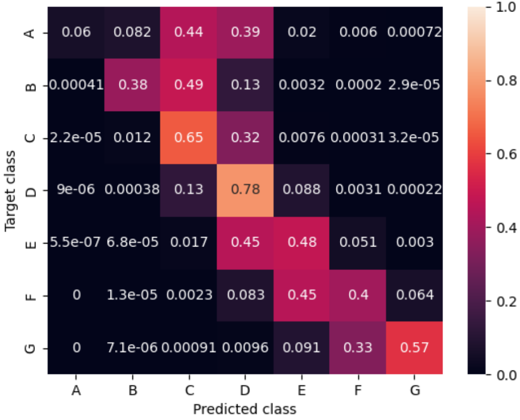

Figure 2: Confusion Matrix for Light Gradient Boosting Machine Model Big with imbalanced dataset.

Imbalanced data

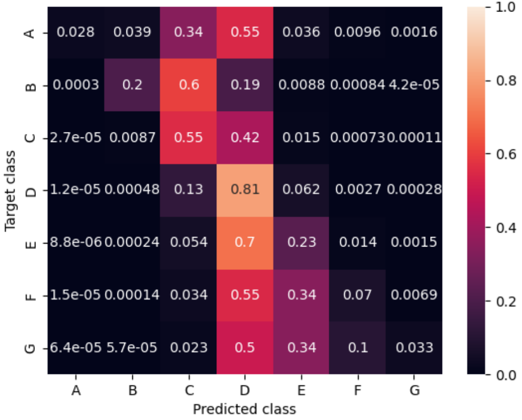

Figure 3: Confusion Matrix for Light Gradient Boosting Machine Small Model with imbalanced dataset.

Imbalanced data

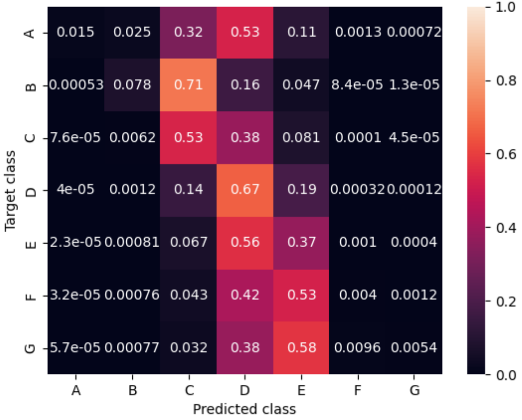

Figure 4: Confusion Matrix for Support Vector Machine Small Model with imbalanced dataset.

Imbalanced data

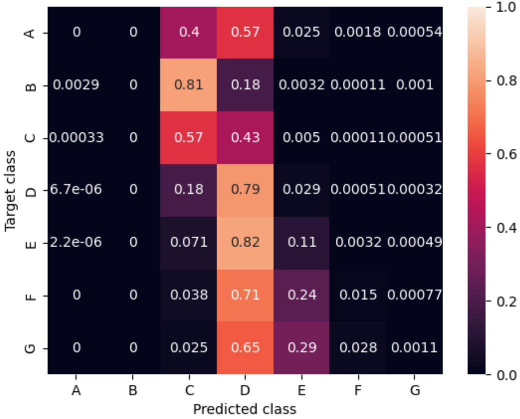

Figure 5: Confusion Matrix for Ordinal Regression Small Model with imbalanced dataset.

Predefined distribution

	D	C	E	F	B	G	A
Predefined	38.9%	22.7%	17.2%	3.9%	16.3%	0.7%	0.3%
Our data	43.8%	29.3%	18.3%	4.7%	2.4%	1.4%	0.1%

Table 4: Distribution of target predefined by the client and in our dataset.

	Light Gradient Boosting Machine Big	Light Gradient Boosting Machine Small
Accuracy	0.632	0.543
Ranked Probability Score	0.045	0.063

Table 5: Accuracy and RPS for models with predefined distribution.

Predefined distribution

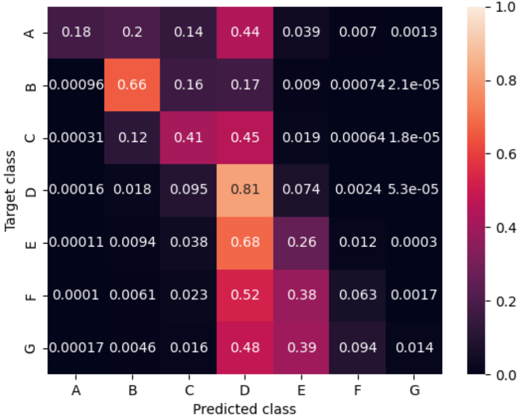

Figure 6: Confusion matrix for Light Gradient Boosting Machine Small trained with the predefined distribution.

Predefined distribution

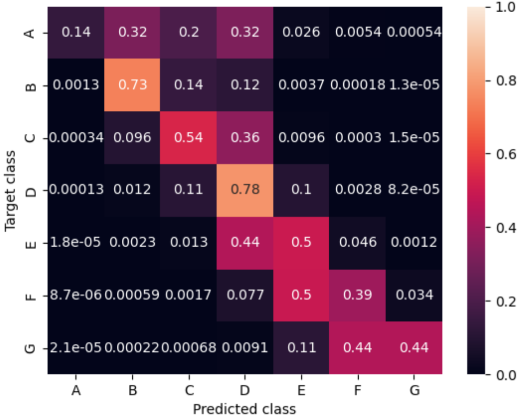

Figure 7: Confusion matrix for Light Gradient Boosting Machine Big trained with the predefined distribution.

Feature Importance

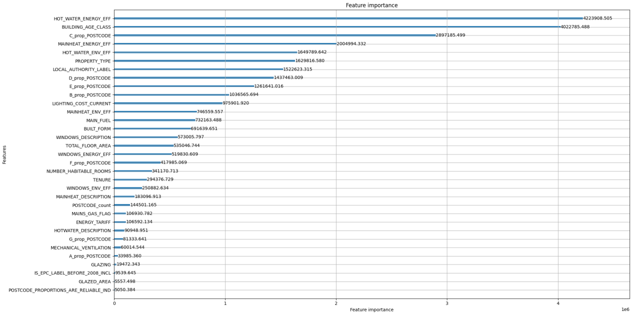

Figure 9: Feature Importance for Light Gradient Boosting Machine Big Model with imbalanced dataset.

Feature Importance

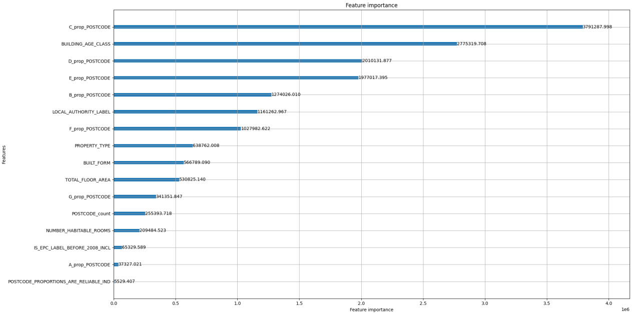

Figure 10: Feature Importance for Light Gradient Boosting Machine Small Model with imbalanced dataset.

Feature Importance

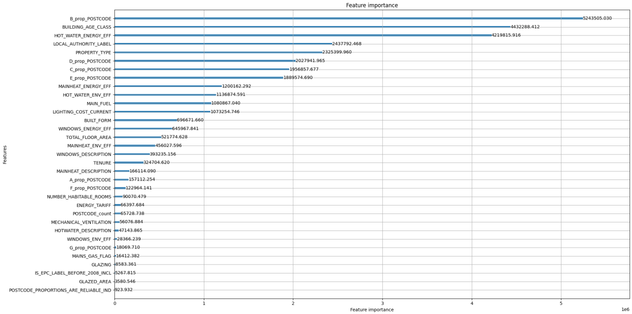

Figure 11: Feature Importance for Light Gradient Boosting Machine Big Model with predefined distribution.

Feature Importance

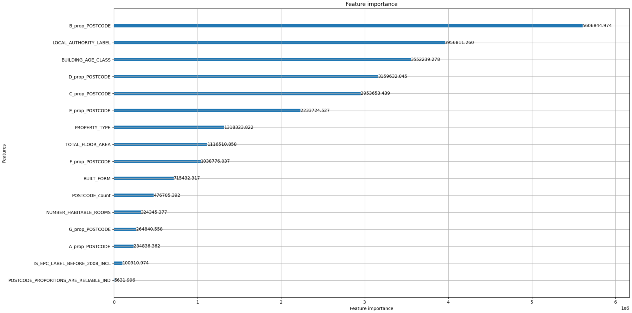

Figure 12: Feature Importance for Light Gradient Boosting Machine Small Model with predefined distribution.

Balanced data

	Light Gradient Boosting Machine Big	Baseline Model Small	Light Gradient Boosting Machine Small	Support Vector Machine Small
A	0.312	0.413	**0.260**	0.303
B	0.096	0.247	**0.088**	0.111
C	0.041	0.088	**0.085**	0.101
D	0.029	**0.027**	0.089	0.083
E	0.061	0.113	0.079	**0.065**
F	0.083	0.259	0.095	**0.094**
G	0.073	0.421	**0.170**	0.177
Weighted Avg	0.043	**0.082**	0.087	0.087

Table 6: Ranked Probability Score for models with balanced dataset.

Light Gradient Boosting Machine Big	Baseline Model Small	Light Gradient Boosting Machine Small	Support Vector Machine Small
0.574	**0.438**	0.422	0.422

Table 7: Accuracy for models with balanced dataset.

Balanced data

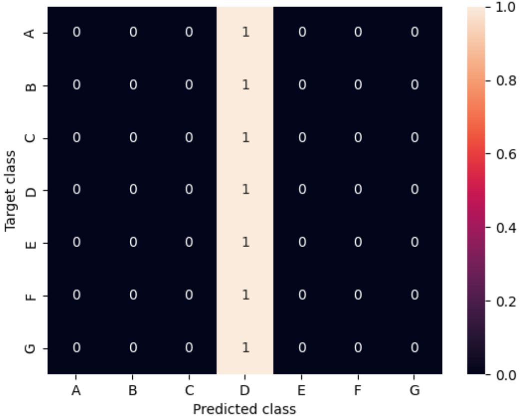

Figure 13: Confusion Matrix for Baseline Small Model (same as with unbalanced data)

Balanced data

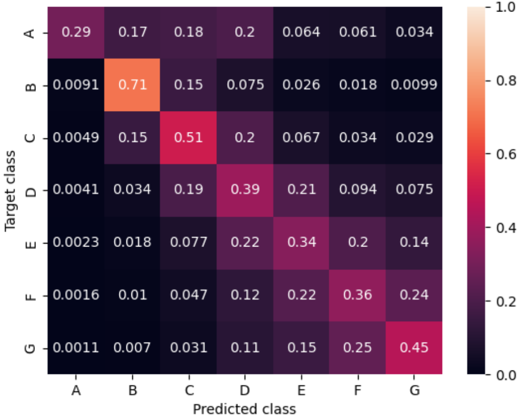

Figure 14: Confusion Matrix for Light Gradient Boosting Machine Small Model with balanced dataset.

Balanced data

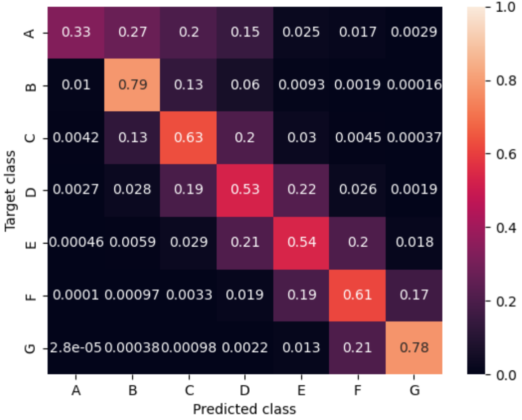

Figure 15: Confusion Matrix for Light Gradient Boosting Machine Model Big with balanced dataset.

Balanced data

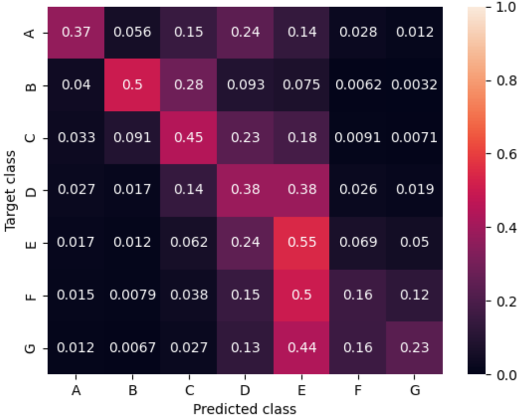

Figure 16: Confusion Matrix for Support Vector Machine Small Model with balanced dataset.

Feature Importance

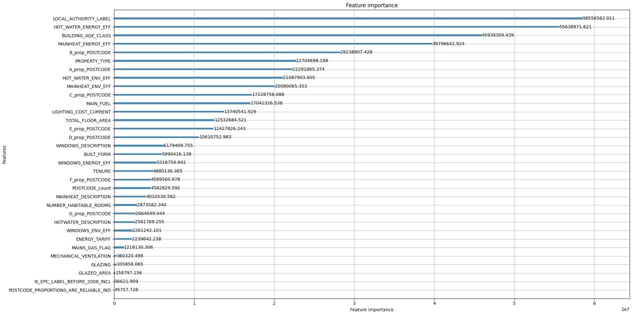

Figure 17: Feature Importance for Light Gradient Boosting Machine Model Big with balanced dataset.

Feature Importance

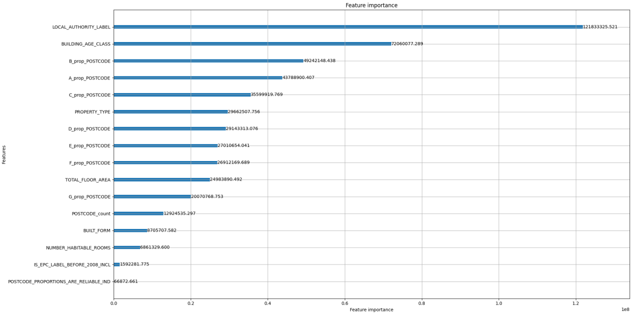

Figure 18: Feature Importance for Light Gradient Boosting Machine Small Model with balanced dataset.

Thank you for your attention

Appendix

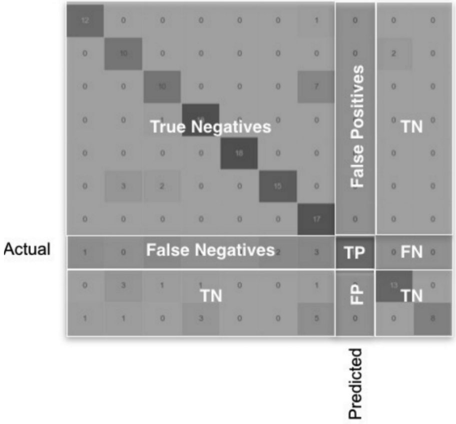

Figure 19: Schematic representation of a confusion matrix and the terms True Positive, True Negative, False Positive and False Negative.

Appendix

$$\mathrm{RPS}((\mathrm{x}^T, y)^T) = \frac{1}{R+1} \sum_{k=0}^{R} \left[\left(\sum_{j=0}^{k} p_j(\mathrm{x}) \right) - \left(\sum_{j=0}^{k} y^{(j)} \right) \right]^2,$$

$$\mathrm{RPS}_i = \frac{1}{|i|} \sum_{((\mathrm{x}_l^T, y_l)^T) \in i} \mathrm{RPS}((\mathrm{x}_l^T, y_l)^T),$$

$$\mathrm{Accuracy} = \frac{\sum_{i=0}^{R} c_{ii}}{\sum_{i,j=0}^{R} c_{ij}},$$

Appendix

- **True Positives** (TP_i): c_{ii},
- **True Negatives** (TN_i): $\sum_{j,k \in \{0,\dots,R\} \setminus \{i\}} c_{jk}$,
- **False Positives** (FP_i): $\sum_{j \in \{0,\dots,R\} \setminus \{i\}} c_{ji}$,
- **False Negatives** (FN_i): $\sum_{j \in \{0,\dots,R\} \setminus \{i\}} c_{ij}$.

Appendix

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i},$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i},$$

$$\text{F1 Score}_i = 2\frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i},$$

where $i = 0, \ldots, R$.