PREDICTING REAL ESTATE PRICES IN MONTREAL USING MACHINE LEARNING METHODS

Maria Kędzierska
Faculty of Mathematics and Information Technology
Warsaw University of Technology
Warsaw, Poland
https://github.com/kaluskam

Szymon Gut
Faculty of Mathematics and Information Technology
Warsaw University of Technology
Warsaw, Poland
https://github.com/Szymon-Gut

Wiktor Jakubowski
Faculty of Mathematics and Information Technology
Warsaw University of Technology
Warsaw, Poland
https://github.com/WJakubowsk

Maciej Orsłowski
Faculty of Mathematics and Information Technology
Warsaw University of Technology
Warsaw, Poland
https://github.com/maciejors

May 12, 2022

ABSTRACT

Machine Learning. The indispensable part of twenty-first century technology's development. Methods of ML are helpful in many fields of our daily live. We decided to utilize ML powerful methods in analyzing real estate market in Montreal. Based on collected data, we want to predict the value of real assets in one of the biggest and most popular cities in Canada.

Keywords Machine Learning · Artificial Intelligence · ML models · housing market · real estate · business · predicting · Data Science

1 Introduction

2 Data

Our dataset consists of six data frames. The core data frame is 'listings.csv', as it contains key attributes for real estates sold in Montreal. During data exploratory analysis we have discovered that some columns contain largely nulls, so we have decided to drop them. Additionally, columns with the same value for all records were also rejected as they did not provide any information. The variable we want to predict is BuyPrice, we observed that some records contained the price per foot or square meter while the rest contained the total price of the property. In addition, we have calculated that the purchase price is mostly correlated with the potential gross income, the number of bathrooms, bedrooms, and rooms.

The next data frames that were analyzed, were 'policeCoord.csv' and 'firestations.csv'. The next data frames that were analyzed were "policeCoord.csv" and "firestations.csv". These data frames only contained the coordinates of the police and fire departments. It seems useful to calculate the distance to the nearest police and fire station for each property.

Next, we examined the file "montreal_hpi.csv" which contained the house price index measures. Due to the significant correlation (greater than 0.99) between the variables, we concluded that only one column should be used from the mentioned dataframe.

In order to add socio-demographic data, we have analyzed 'sociodemo.csv'. A data frame with information about population, nearby schools, average income, etc.

The last data frame was 'extra_data.csv' which provided additional information for some properties in 'listings.csv', unfortunately this frame was contained vast majority of null values. However, some columns, such as LivingArea, found themselves truly relevant.

3 Selected models

After exact EDA (exploratory data analysis) we started next step which was building and testing different models. We have built several RandomForest, XGBoost, LightGBM and CatBoost models. Our models were trained using Grid Search and Random Search. Then were selected only the best models based on RMSE. For each alghorithm mentioned before we have selected five models with different hyperparameters which have reached the best scores. Then we have tried PCA to reduce dimension of the dataset and this significantly improved scores of our models. Below is a table with the best scores on test data sets from the four selected algorithms.

RMSE	**RandomForest**	**XGBoost**	**LightGBM**	**CatBoost**
PCA	87 740.23	360.06	74 396.47	46 023.39
WITHOUT PCA	334 091.74	228890.90	282 358.90	246 424.10

All hyperparameters were set to avoid overfitting and underfitting so models presented above have similar scores on training and test data sets. Then we selected the best algorithm for our models, which, as we can see in the table above, turned out to be XGBoost.

Below we can see the best 3 models built on XGboost's algorithm and their hyperparameters

Model	**RMSE train**	**RMSE test**	**booster**	**eta**	**max_depth**	**gamma**	**min_child_weight**	**subsample**	**colsample_bytree**
xgboost 1	101.90	360.07	gbtree	0.05	7	0	1	1	1
xgboost 2	430.97	881.18	gbtree	0.09	8	4	1	1	1
xgboost 3	808.91	1 477.11	gbtree	0.12	6	6	1	1	1

Next, we created residual plots to check how our predictions looked like. What is really satisfying is that almost all predictions are very accurate. Residual plots are presented below.

Figure 1: First model

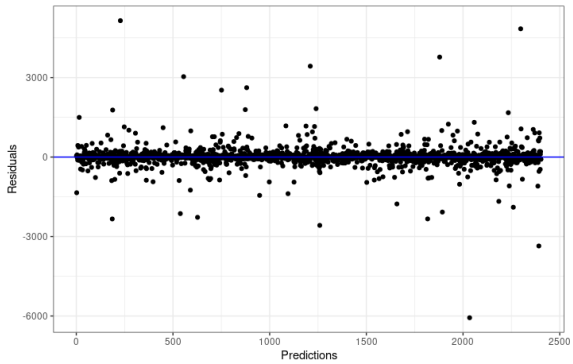


Figure 2: Second model

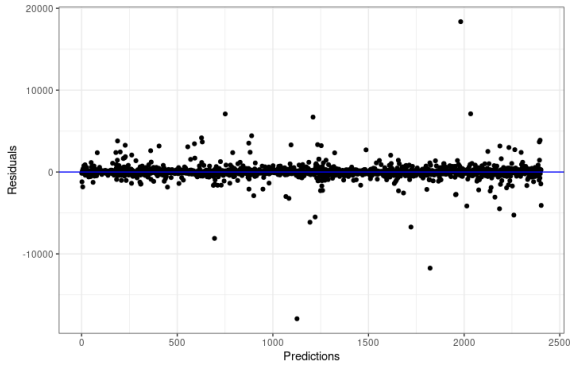
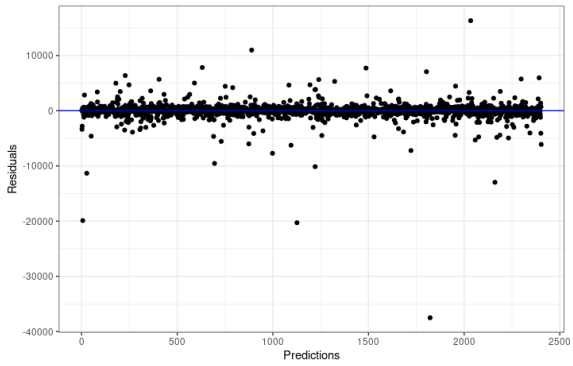


Figure 3: Third model



So as we can see these three XGBoost models were very accurate and reached very low RMSE. But how about the other metrics? That was our next step in choosing accurate model. We have measured the predictions using MAE, R^2 and the mentioned before RMSE. In the following formulas, y_i stands for real value of i-th observation in dataset and $\hat{y}_i$ stands for the model's prediction of this value.

MAE is defined as $\sum_{i=1}^{n} |y_i - \hat{y}_i|$,

R^2 is defined as $\frac{\sum_{i=1}^{n}(\hat{y_i} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})}$,

and the RMSE is defined as $\sqrt{\frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}}$

So now we can present our results:

Model name	**RMSE test**	**MAE test**	**R2**
xgboost1	360.07	135.74	0.9998625
xgboost2	881.18	289.28	0.9995856
xgboost3	1 477.11	520.25	0.9978164

All three models achieve very good metrics and will also be further analyzed using XAI methods. But there is one change that must be performed. As we have used the PCA method to reduce dimension of our data sets we have lost the interpretability of our data. So we need to decide what is more important: increasing the accuracy of the prediction or preserving the interpretability and an insight into the influence of individual variables on the predictions. In the next steps, where we are analyzing our models using XAI methods, we are using the data set without PCA.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis.

Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 3.

3.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{1}$$

3.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

4 Examples of citations, figures, tables, references

4.1 Citations

Citations use `natbib`. The documentation may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Here is an example usage of the two main commands (`citet` and `citep`): Some people thought a thing [Kour and Saabne, 2014a, Hadash et al., 2018] but other people thought something else [Kour and Saabne, 2014b]. Many people have speculated that if we knew exactly why Kour and Saabne [2014b] thought this...

4.2 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 4. Here is how you add footnotes. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

[1] Sample of the first footnote.

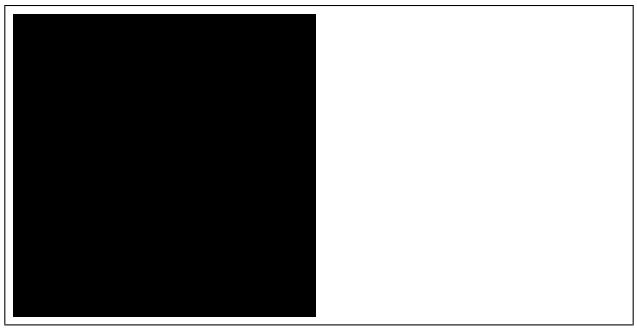


Figure 4: Sample figure caption.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

4.3 Tables

See awesome Table 1.

The documentation for `booktabs` ('Publication quality tables in LaTeX') is available from:

`https://www.ctan.org/pkg/booktabs`

4.4 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

References

George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE, 2014a.

Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE, 2014b. doi:10.1109/SOCPAR.2014.7008025.