



## **HOUSING PRICE PROJECT REPORT**

Submitted by:  
M.S.JISHNU

## ACKNOWLEDGEMENTS

I would like to express our deep and sincere gratitude to FLIP ROBO for giving us the opportunity to do this project. As a great bridge between academic and industry, this program educated us how to perform theoretical methodology in real life.

We would like to express our sincere thankfulness to Mohd Kashif for the continuous support of our research, for their patience, enthusiasm, motivation and immense knowledge. You kept reminding us of the importance of collaboration, which ensured that the whole project proceeded smoothly.

## 1. INTRODUCTION

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors.

Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry. Notably, this has been done in Zillow's Zestimate [4] and Kaggle's competitions on housing prices [2].

In some cases, non-traditional variables have proved to be useful predictors of real estate trends. For example, in [3] it is observed that Seattle apartments close to specialty food stores such as Whole Foods experienced a higher increase in value than average.

This project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focused on assessment value for residential properties in Calgary between 2017-2020 based on data from [1]. The aim of our project was to build a predictive model for change in house prices in the year 2021 based on certain time and geography dependent variables.

The main steps in our research were the following.

- **Exploratory Data Analysis (EDA).**

By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting.

- **Feature Selection**

In order to avoid overfitting issues, we select 20 (according to PCA [12])

As part of the EDA, we first looked at the mean percent change of the housing prices from 2017-2020 for each FSA whose data is given. Figure 2 suggests that on

an average there was positive change in prices in the Year 2018. In order to analyse our features more carefully, we also looked at the correlation of various features of the houses.



FIGURE 3. Correlation of Features

Figure 2 gives us an insight on how parameters are correlated with each other.

### 3. METHODOLOGY

**3.1. Feature selection.** Our data has 36 features in total. If we use all of them in our prediction model, the model will have a risk of overfitting. Therefore, we decide to remove some unimportant features. We choose a dimensionality reduction algorithm called Principal Component Analysis (PCA) as the method to estimate how many components are needed to describe the data. The optimal number of features for the prediction can be determined by looking at the cumulative explained variance ratio as a function of the number of components.

This curve quantifies how much of the total, 36-dimensional variance is contained within the first  $n$  components. For example, we see that with the digits the first 10 components contain approximately 90% of the variance, while you need around 25 components to describe close to 100% of the variance. Here we see that our

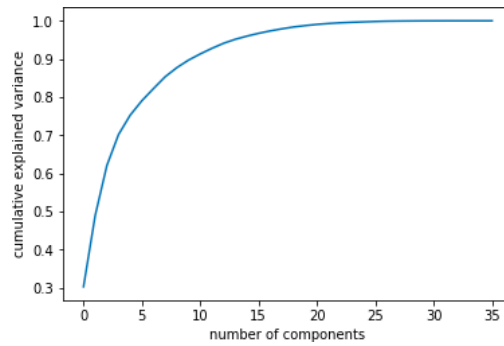


FIGURE 4. PCA Analysis

five-dimensional projection loses a lot of information (as measured by the explained variance) and that we would need about 20 components to retain 98% of the variance.

By using five different feature selection methods: ANOVA, LASSO, ELASTIC NET, FORWARD FEATURE SELECTION, BACKWORD FEATURE SELECTION, we were able to select 20 features out of the initial 36 features as a result of overlapping features that we observed in each feature selection algorithm. Those 20 features are: saf4; saf5; mr5y; Inflation; pop1; pop2; inc3; own3; lab1; walk score comm; Age; saf2; saf3; pop3; pop4; inc1; inc2; own2; lab2; vacancy rate.

**3.2. Percent Change price prediction.** The percent change price can be divided into four different groups:  $[-0.12, -0.06)$ ,  $[-0.06, 0)$ ,  $[0, 0.06)$  and  $[0.06, 0.12]$ . In this section, we are going to consider our problem as a classification problem.

Based on the selected features, we applied three different Machine Learning algorithms: Decision Tree, Random Forest and XGBoost, on the training data and then used the testing data to check the accuracy, which equals to the number of samples that predicted in the right group divides the total sample size of our testing data. Here is the table of the accuracy rate:

Method	Accuracy Rate
Decision Tree	66.8%
Random Forest (with 1000 estimators)	68.1%
XGBoost	69.7%

Since XGBoost model is interpretable and performs best on the accuracy rate, we use it as our prediction model.

The above tree plot gives an example on how does an XGBoost model arrive at its final decision. This plot also shows the conditions on the node that splits the tree.

After getting an XGBoost model, we can examine the importance of each feature within the model by counting the number of times each feature is split on across

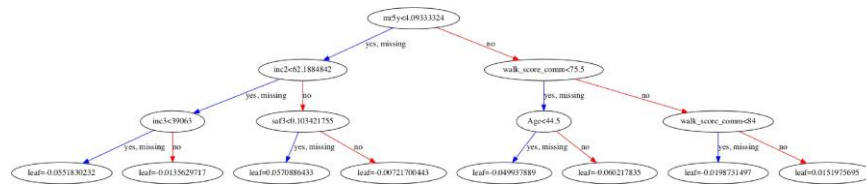


FIGURE 5. XGboost Decision Tree

all boosting trees in the model. The order of the importance of different features is plotted as a bar graph:

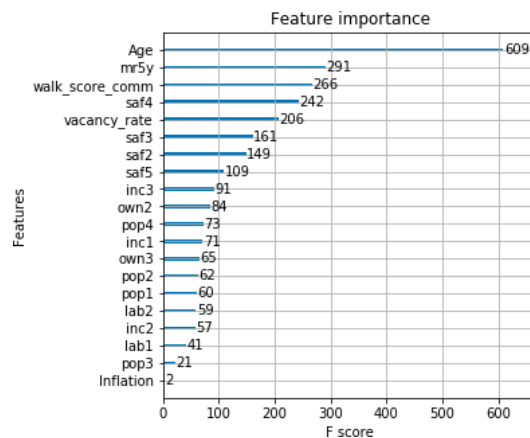


FIGURE 6. Feature Importance Bar Plot

From this plot, we can see that *Age* has the highest importance and *Inflation* has the lowest importance.

Next, we use the confusion matrix to help us visualize the performance of this XGBoost model:

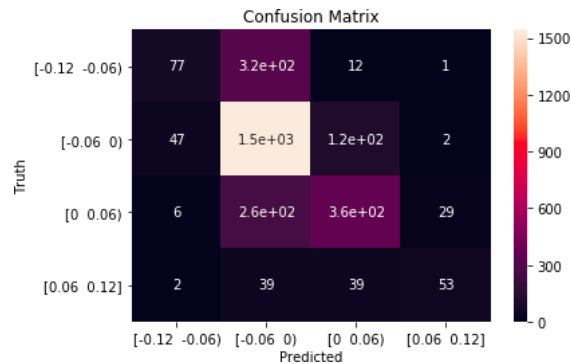


FIGURE 7. Confusion Matrix Results

In the above matrix, each row represents the instances in an actual group while each column represents the instances in a predicted group. It is every easy to see how many samples are mislabeled by this model. Take group  $[0.06, 0.12]$  as an example, the actual size of that group is 133, 53 of them are predicted correctly in the group  $[0.06, 0.12]$  while 80 of them are mislabeled in the wrong groups: 39 cases are mislabeled in group  $[0, 0.06)$ ; 39 cases are mislabeled in group  $[-0.06, 0)$  and 2 cases are mislabeled in group  $[-0.12, -0.06)$ . The above matrix shows that most cases in group  $[-0.06, 0)$  can be predicted correctly by this XGBoost model, but the mislabeling rate for other three groups is not low, especially for group  $[-0.12, -0.06)$ . Therefore, our next step is to find the main reasons for those mislabeling cases.

#### 4. EXPLORATION OF REASONS FOR MISCLASSIFICATION IN MODEL

We focus on finding the reasons why some houses that are supposed to appear in group  $[-0.12, -0.06)$  are in group  $[-0.06, 0)$  and houses supposed to appear in  $[0, 0.06)$  appear in  $[-0.06, 0)$ . Thus, we could check datum of significant factors and find out why misclassification occur. Why this is important—we want to find out why the predicted percentage change of price for some houses is exceptionally low or high, which is important to basically every stakeholder. The following is a screen shot of our dashboard and the dots on the map are some selected points from our table.

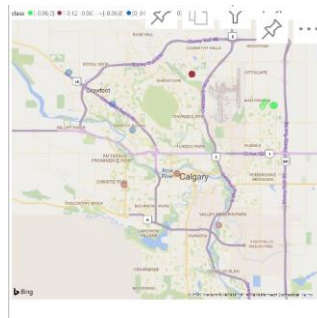


FIGURE 8. Some houses with exceptionally high/low percent change

For further information of other properties, please refer to the links in footnotes<sup>12</sup>. After that, we count the frequency of each significant features that appears, and we get the following graph

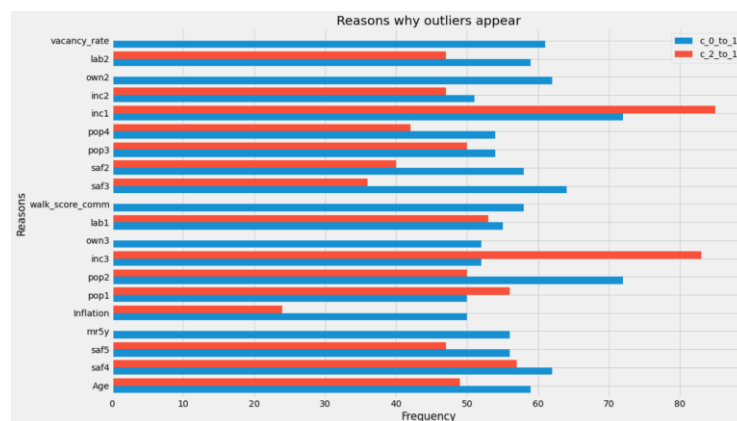


FIGURE 9. Frequency of Significant Features in Outliers

From figure 9, we could see that inc1 and pop1 are the most important factors making some houses that are supposed to appear in group  $[-0.12, -0.06)$  are in group  $[-0.06, 0)$ . Also, inc1 and inc3 are the most important factors making houses supposed to appear in  $[0, 0.06)$  appear in  $[-0.06, 0)$ .

## 5. COMMUNICATING OUR RESULTS

Given that our project was motivated by practical interest to stakeholders, we aim to publicly deploy a dashboard presenting our main results. This part of the project is still work in progress; we are currently working on a prototype in Power



BI (see Figure 10), and we will then build the dashboard using Plotly Dash since this allows for easy public deployment.

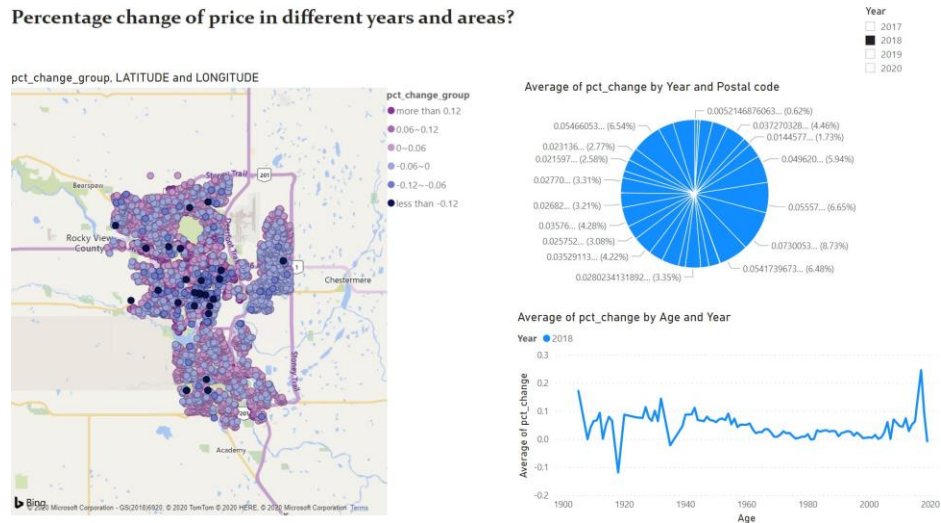


FIGURE 10. Dashboard prototype

Much of our data can be naturally viewed on a map, and by doing this many properties of the data can be seen easily. Specifically we place the set of data points on a map of Calgary using the coordinates of each house. By colouring these points according to the percent change in price of the corresponding house, the user can then visually identify geographical patterns.

The dashboard<sup>3</sup> will be interactive, giving the user the ability to view our results from various perspectives. For example, the user will be able to select a geography-dependent variable of interest to them, and to view the map colour-coded according to this feature with the data points overlaid. Different stakeholders may be interested in different variables, and these interactive features allow each user to choose how they visualise the data.

In a separate section of the dashboard, we will have our predictive model. In this section, the user test our model on new data points and the model will output a prediction for the percent price change of that house in the year 2021. This allows the user to explore how house prices would react in various potential scenarios, and we believe this could be helpful for future decision making.

## 6. SUMMARY

By analysing historical data for house prices in Calgary along with various relevant features, we established some interesting patterns and trends. Using machine

learning techniques, we were then able to identify a subset of the original features that are in a sense sufficient to describe our data.

Having selected the most important features, we then trained an XGBoost model for change in house price prediction, which classified samples into one of four categories. This model gave an accuracy rate of 68.7 on a test set that we had kept separate during development. This model can therefore be used to predict, for example, which type of house within Calgary is likely to increase and decrease in price in the year 2021 based on various scenarios.

## REFERENCES

1. *data source*, [https://raw.githubusercontent.com/shughestr/PIMS\\_2020\\_Real\\_Estate\\_data/master/sample\\_clean.csv](https://raw.githubusercontent.com/shughestr/PIMS_2020_Real_Estate_data/master/sample_clean.csv).
2. *Kaggle competition*, <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>.
3. *Mckinsey report*, <https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate>.
4. *Zillow zestimate*, <https://www.zillow.com/blog/zestimate-updates-230614/>.
5. Kam C Chan, Patric H Hendershott, and Anthony B Sanders, *Risk and return on real estate: evidence from equity reits*, Real Estate Economics **18** (1990), no. 4, 431–452.
6. T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (2016, August), 785–794.
7. B. de Ville, *Decision trees*, Wiley Interdisciplinary Reviews: Computational Statistics **5** (2013), no. 6, 448–455.
8. T. G. Dietterich, *An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization*, Machine learning **40** (2000), no. 2, 139–157.
9. R. A. Fisher, *Statistical methods for research workers*, In Breakthroughs in statistics (1992), 66–70.
10. Franz Fuerst and George Matysiak, *Analysing the performance of nonlisted real estate funds: a panel data analysis*, Applied Economics **45** (2013), no. 14, 1777–1788.
11. Richard J Herring and Susan M Wachter, *Real estate booms and banking busts: An international perspective*, The Wharton School Research Paper (1999), no. 99-27.

