**STA2453 Project Report**

**Summary (one paragraph) - Do after report**

**Introduction (~2 Pages)** - (Shea)

* Purpose
* Scope
* Background

**Methods (~3 Pages)**

* Data Scraping - (Shea)
* Data Preprocessing/Cleaning （Siyuan）

Describe the data

* Exploratory Analysis (plots) （Siyuan / Ida）
* Models （Siyuan / Ida）

**Results (~3 Pages)**

* Linear Regression （Siyuan）
* Model Comparison(LR vs LASSO vs Random Forest) - (Ida)
* Spatial Analysis - (Shea)

**Discussion/Conclusion (~2 Pages)** - (Ida)

* Summarize result
* Limitations

More data

* Further work

Categorical encoding technique: target encoding

**Reference**

**Appendix**

**1 Introduction**

**2 Methods**

**2.1 Data Scraping**

**2.2 Data Preprocessing/Cleaning**

**2.3 Exploratory Data Analysis**

The distribution of the condo price per $1000 is plotted below, which shows a positive skewness with a long right tail. After performing a log transformation to the price, the distribution looks much more normal. This indicates that a log transformation may help improve the prediction performance.

**Histogram

Description automatically generated with medium confidenceChart, histogram

Description automatically generated**

A boxplot is drawn to show the relationship between the number of bedrooms and price. We could tell from the plot that as the number of bedrooms increases, the average property price increases gradually when the number of bedrooms is between 1 and 4. There are many outliers for the properties with two or three bedrooms. It indicates the price could be affected by other indexes of the apartment. When there are 5 and 6 bedrooms, the average price is lower than the apartments with 4 bedrooms. The reason is because the users input the living room and kitchen as a bedroom. This could mislead the prediction. Thus, we exclude those records in the modeling.

**Chart, box and whisker chart

Description automatically generated**

The correlation heatmap is also plotted to explore the correlation between variables, where light color means high correlation. When looking at the response price and predictor variables, the plot indicates the number of bathrooms, the number of bathrooms, and maintenance fee are the important features to predict price. Multicollinearity can also be detected by looking at the relationship between predictor variables. Most of the predictors do not show high correlation except waterfront community and the first three digits postcode of M5V. Thus, we remove the postcode M5V and only keep the community information when training the model.

**Graphical user interface

Description automatically generated with low confidence**

**2.4 Models**

2.4.1 Linear Regression

2.4.2 LASSO Regression

The least absolute shrinkage and selection operator(LASSO) regression is a regularized linear regression algorithm, which was first introduced by Tibshirani in 1996. It regularizes the linear regression by adding a penalty term to the loss function. The loss function of LASSO regression shows as follows:[formula]

Here the penalty term is in absolute form, which is also known as the L1 penalty. It penalizes the insignificant features by pushing those coefficients to 0. Since LASSO allows coefficient values equal to 0, it also performs an automated feature selection. It also benefits in avoiding overfitting.

2.4.3 Random Forest

Random Forest is a popular machine learning algorithm, which consists of multiple decision trees. The key component is splitting on features until getting a reasonable prediction score​. It is known as supervised ensemble learning, which combines the  prediction of multiple models together to make more accurate predictions. Each tree in a random forest is in parallel. Compared with a single tree, it can avoid high variance and achieve strong predictive power.

**2.4 Categorical Variable Encoding**

Since the realtor dataset includes many categorical variables, categorical variables encoding techniques are required before fitting the models. There are two main types of categorical variables: ordinal and nominal data. Ordinal data is categorical data with order, which can be handled by label encoding. Nominal data is categorical data without order. The most common solution to deal with nominal data is one-hot key encoding. That is creating a new dummy variable​ for each level of a categorical feature. However, many categorical variables in the dataset include dozens of levels such as community and management company. If directly applying one-hot encoding to categorical variables, there is risk of sparsity and computational inefficiency. Eventually, it will deteriorate the overall performance of tree-based models. Thus, we combine the one-hot encoding with manual encoding, which selects top popular levels in each categorical variable to reduce the overall number of variables.

**3 Results**

**3.1 Linear Regression**

**3.2 Model Comparison**

We also compare the performance of linear regression with other more sophisticated models. There are 5 models in total: linear regression, linear regression with log transformation, LASSO regression, random forest without feature selection, and random forest with feature selection. The features of random forest with feature selection are selected by the top 10 most important features from the original random forest.  All the models are trained and evaluated with 75/25 train/test split using R square and root mean-squared-error (RMSE). The results are shown in the table below. The best score is bolded.

Table

Description automatically generated

From the table, we could tell that the random forest model without feature selection best fits the training set with lowest R square (0.823) and RMSE(176694). It is followed by the random forest with feature selection. However, random forest models gain low scores in the test set, which do not even achieve the predictive power as a simple linear regression model. This indicates the overfitting problem. The reason is the dataset includes many categorical variables, and random forest models are biased to features with more levels​. We could also find that the random forest model with feature selection does not have effective feature selection compared with original random forest. It may ignore other importance levels in categorical features by using one-hot key encoding. For the test set, linear regression with log transformation outperforms other models, which shows great prediction performance with 0.639 R square. As we expected, log transformation helps linear regression achieve a better performance in both training and test set by increasing normality and reducing the variability of data. LASSO regression does not show strong performance as simple linear regression. That is because the scale of data is large, which leads to large coefficients. LASSO does not benefit from pushing coefficients to zero and perform the automated feature selection.

Graphical user interface

Description automatically generated

The top 15 feature importance of the random forest model is shown above. We found Maintenance fee, number of bathrooms and bedrooms, and Annex & Waterfront community play an important role in predicting the condo price.

**3.3 Spatial Analysis**

**4 Discussion/Conclusion**

While the linear regression model with log transformation shows great prediction performance among models, there are still limitations in our methods. In the web scraping part, even though we could extract condo information with the help of webdriver Selenium and BeautifulSoup, there is still manual intervention required because of the additional security check warning. This limits the data volume extracted. There are some packages which could bypass the captcha restriction, such as Anticaptcha and 2Captcha. However, since both services involve real humans to deal with the challenge, it will be costly.

Moreover, since most of the machine learning models in Python could not handle categorical variables directly, we applied one-hot key encoding to convert categorical variables to numeric. Whereas, it could lead to exponentially increased numbers of variables when there are many levels in categorical variables, which will cause the overfitting problem and decrease the predictive accuracy, especially for tree-based models. Hash encoding is a potential solution. Instead of introducing a new dimension for each level, it uses the hashing algorithms to transform data into lesser features.

In algorithm selection, we have tried both linear regression and sophisticated models like random forest. However, the random forest models do not show great performance as we expected since the overfitting problem. That is because of the characteristics of the dataset, which includes many categorical variables with many levels. Therefore, we suggest that data scientists should never rely too much on sophisticated models. Models should be selected based on the problem and dataset.

**Conclusion**

In this report, we applied multiple algorithms to predict the Toronto condo price. The linear regression model with log transformation shows the best prediction performance. Among all features, maintenance fee, number of bedrooms, number of bathrooms, and community play the most important roles when predicting the price.

**Reference**

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". *Journal of the Royal Statistical Society*. Series B (methodological). Wiley. 58(1): 267–88. [JSTOR](https://en.wikipedia.org/wiki/JSTOR_(identifier)) [2346178](https://www.jstor.org/stable/2346178)