Project Plan and EDA

STAT5003 Group 8

Yudan Wu, SID:510498173 Sheng Chang, SID:450212420
Xingyu Chen, SID:510559289 Long Cheng, SID:510097921
Stephen Ieong, SID:510049458

Semester 2, 2021

1 Classification Statement

The classification method of our project is to predict the real estate price in different regions in Russia based on the different features provided [1]. With the collapse of the Soviet Union, Russia's economic model was transformed to private commerce system, so housing transactions and rentals became important components of Russia's economic growth. The prediction model could be helpful for individuals to determine the overall future price changes of an estate property, and for tenants to ensure the following behavior corresponding to trends. For Russian Real Estate Department, they can utilize this data set to determine the future price trends of a particular property and ensure the next macro development of the area where the property is located, finally combine all areas for searching direction of real estate development in future.

2 Dataset Description

The dataset keeps 13 variables as 13 columns with 500 million instances of different house estates around the whole Russia, the total dataset comprises four sorts of data types, and "Price" is the desired label variable, the new variable "range_price" has been introduced to categorise the original "price" variable into nine interval groups for better classification, and the "area" variable would be isolated as "geo_lat" and "geo_lon" would exert a better effect. We chose one of the most representative regions of Russia's emerging real estate economy as our object of observation, with 27827 instances. The nine interval groups were created by the code below.

2.1 Data Types

The "Region" represents 85 areas in Russia whilst the "building_type" illustrates five different types of building facades from number 1 to 5, it means Other, panel, Monolithic, Brick, Blocky and Wooden in order. The "object_type" means the ownership status, 1 means preowned while 2 means new. The "Date" variable is an ordinal data contains the specific announcement date of different estates from 2018 to 2021, "Time" means the particular time in each day of an estate information update. The "geo_lat" and "geo_lon" means the Latitude and Longitude of this building respectively. The "level" means the floor of the apartment located, and "levels" means the highest floor inside this building. The "rooms" means the number of living rooms, when it equals -1 illustrating it is a studio apartment. The "area" means the total area of this apartment, and the "kitchen_area" means the kitchen area in square meters. Moreover, the "Price" is the label variable which represents the price of this building in rubles.

3 Graphical Description

3.1 Data Cleaning

For outlier investigation, the project would utilise the price attribute to create boxplot visualisation. The boxplot indicated a data point with a minimal price and a certain amount of data points with significant prices around 1000 million roubles. These data would be considered as outliers and would be removed by code. The boxplot was presented with the histogram in section 3.2.

3.2 Histogram

Based on feature price, the project attempts to generate a new variable, which was the range_price, as the objective attribute. Before that, the project uses visualisation for understanding the distribution of price. Therefore, a histogram was created to show that region 5368 has a massive amount of data points were between 200 million and 400 million roubles in Russia. The code is presented below. **(Click on the figure to zoom in, click again to zoom out)**

3.3 Correlation Matrix

A correlation matrix is an excellent method to display the relationship between each feature. Based on the project's correlation matrix, range_price (objective variable) has a relatively higher correlation with price, rooms and levels; as high contributed explanatory variables, the project will focus on them while building and training classification models. Instead of a higher correlation between levels and level, area and room, the correlation between explanatory variables are relatively low. The package corrplot was utilised for creating correlation matrix [2].The code is presented below. **(Click on the figure to zoom in, click again to zoom out)**

4 Potential Challenges

4.1 Large Dataset

Due to the large size (27827 rows) of the data set, the project Due to the large size (27827 rows) of the data set, the project would encounter enormous computational load on learning and training the models. For conquering this issue, the project would utilise parallel computing methods. R has provided built-in packages for parallel computing [3], [4].

4.2 Overfitting

Overfitting is a common and severe issue which would occur during learning and training the models. Overfitting indicated that the training model perfectly fits the training set and could not generalise future examples properly. The solution of the project would be applying regularisation and penalty on each classifier.

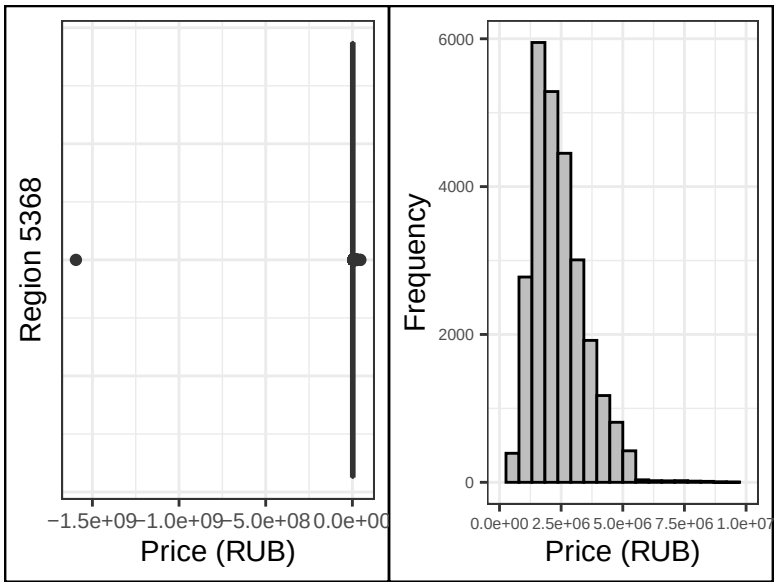

Figure 1: Boxplot and histogram of region 5368

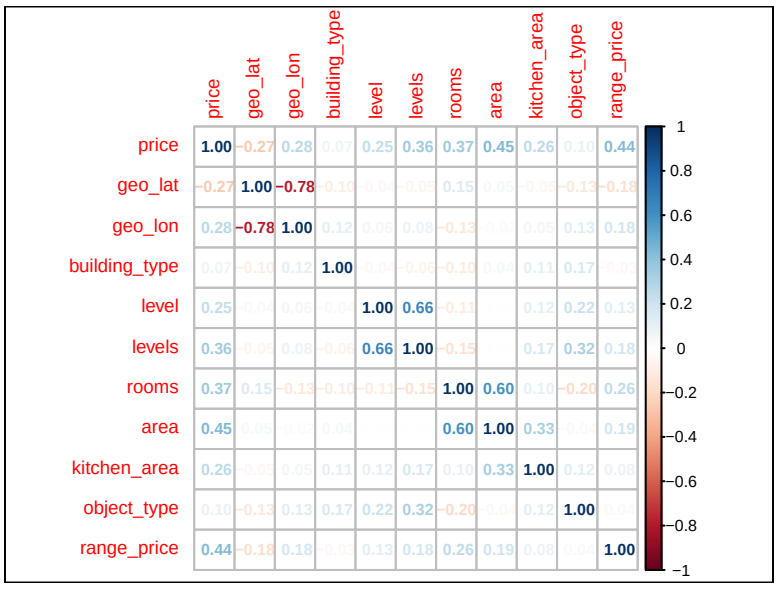

Figure 2: Correlation Matrix

4.3 Imbalance of Class

As the objective variable, range_price has nine classes. Due to the difference in price distribution, each class of price range was not equally distributed. If class imbalance causes inconvenience in learning models and predicting outcomes, the solution would be adjusting and rebalancing the size per class. The nine classes's code and tuples distribution would be given in the code below.

```
##
##    0    1    2    3    4    5    6    7    8
## 1315 1109 4241 5867 4665 3994 2531 1638 1038
```

5 Performence Metrics

5.1 Classification Report

Classification report was an excellent method for determining the performance of the classifiers. The report would calculate and present the accuracy score, precision, recall and F1 score. The equations of the four scores were presented below.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{1}$$

$$Precision(P) = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall(R) = \frac{TP}{(TP + FN)} = \frac{TP}{P} \tag{3}$$

$$F_1 = \frac{2 \times P \times R}{(P + R)} \tag{4}$$

5.2 Confusion Matrix

Confusion matrix illustrated the allocation of true positive, false positive, true negative and false negative of the prediction.

6 Models Chosen

6.1 Logistic Regression

Logistic regression is known as logistic regression analysis. It is a machine learning method used to solve binary classification problems, which is used to predict probability or determine the belonging category.

6.2 K-Nearest Neighbors

The K-nearest neighbour is also called the KNN algorithm, which is supervised machine learning. The basic principle of KNN is to assign the test data to its belonging category. Moreover, the KNN algorithm has different distance measurement methods such as Euclidean distance, Manhattan distance, etc.

6.3 Linear Discriminant Analysis (LDA)

The LDA algorithm is a supervised learning dimensionality reduction method, which means that each sample of its data set has a category output. The basic idea is to lower the data in low dimensions and expect each category of data's projection points to be close to each other. However, the distance between different categories should be as large as possible.

6.4 Support Vector Machine (SVM)

The SVM algorithm is a supervised learning model and is suitable for both binary classification and regression analysis. When giving a training data set, each training set is marked as belonging to one or the other of the two categories.

7 Schedule

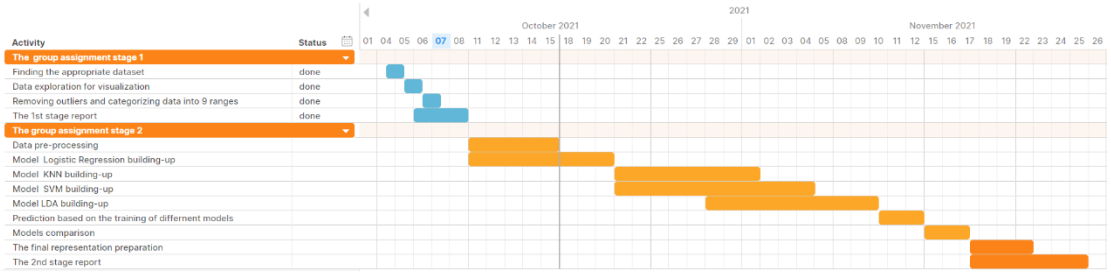

Figure 3: Gantt Chart

Figure 3 illustrated the work content division before the deadline, as the 1st stage content has been done so far, and the following part would be generally divided into data pre-processing, models building, models comparison and final presentation and literature exhibition, in corresponding to the division, a reasonable time distribution has been assigned, in addition an

appropriate extension is considered into it for any inevitable changes. **(Click on the figure to zoom in, click again to zoom out)**

8 Reference

[1] Daniilak, "Russia real estate 2018-2021." Kaggle; Available at https://www.kaggle.com/mrdaniilak/russia-real-estate-20182021, 2021.

[2] T. Wei and V. Simko, *R package 'corrplot': Visualization of a correlation matrix*. 2021.

[3] S. Urbanek and L. Tierney, *R package 'parallel': Support for parallel computation in r*. 2020.

[4] M. Wallig and S. Weston, *R package 'doParallel': Foreach parallel adaptor for the 'parallel' package*. 2020.