**ALY6040 90248 Data Mining Applications SEC 01 Summer 2023 CPS [BOS-D-HY]**

**Module 3 Assignment — Technique Practice**

A black and white logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Hanchan Xu  [xu.hanc@northeastern.edu](mailto:xu.hanc@northeastern.edu)

Zihan Ma [ma.zihan1@northeastern.edu](mailto:ma.zihan1@northeastern.edu)

**Instructor**

 Prof. Kasun Samarasinghe

**Date**

08/02/2023

**Technique Practice For Singapore Airbnb Price Zone**

**Part 1**

**Introduction**

This report presents the data analysis process performed on the listings.csv dataset. The data features various attributes of different listings, which are classified into three categories based on their price: low, middle, and high.

This part of the analysis's primary objective is to predict a listing's price category using a Support Vector Machine (SVM) model.

**Data Preparation**

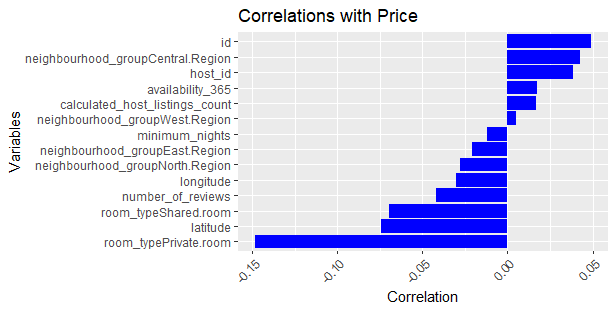
The raw data was preprocessed before feeding into the SVM model. This process includes the following steps:

1. **Classification of price into three groups (low, middle, high)**: Using the 1/3 and 2/3 quantiles of the price as thresholds, we divided the listings into three groups. The price thresholds were found to be 85 and 164, respectively.
2. **Handling missing values**: We filled in missing values in the 'reviews\_per\_month' column with 0.
3. **One-hot encoding**: The categorical variables 'neighbourhood\_group' and 'room\_type' were one-hot encoded.
4. **Data Scaling**: All numeric columns were scaled to ensure that the SVM model was not unduly influenced by the measurement units.

**Exploratory Data Analysis**

A diagram of a graph

Description automatically generated with medium confidence

By looking at the correlation analysis we found that price doesn't have a high enough correlation to another feature, which means the model will be difficult to use these features to predict the price. 

The correlation between the numeric variables and the target variable (price) was computed to understand their relationship. The 'room\_typePrivate.room' and 'latitude' variables showed the highest negative correlation with the price, while the 'room\_typeShared.room' and 'id' variables showed the highest positive correlation. I have tried improve the model by delete the vireable such as `neighbourhood\_groupWest.Region`, but the result didn’t change obviously.

In the end, the column price, id, and host\_id have been deleted, because they are not meaningful to the model.

**Model Training and Evaluation**

We used the SVM model for the classification task and divided the data into a training set (70% of the data) and a test set (30% of the data). The initial SVM model achieved an accuracy of 64.69% on the test set.



Hyperparameter tuning was performed to optimize the SVM model's performance. We used a random search approach over the parameter space to find the best values for 'C' and 'sigma'. The tuned model achieved an improved accuracy of 70.21%.



Following the new parameter get from random optimiztion.

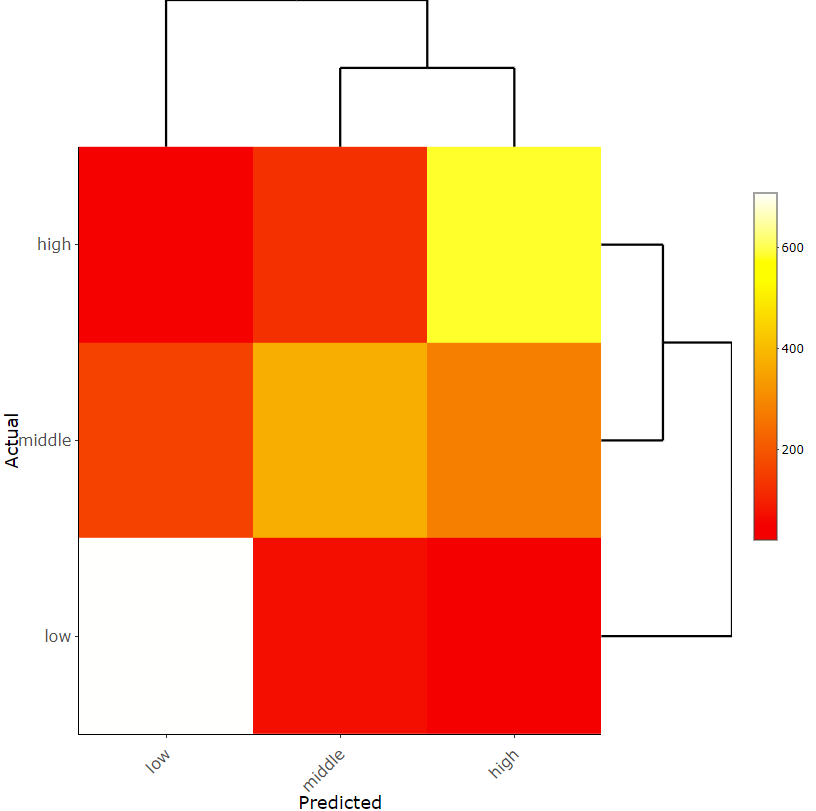
A black and white image of numbers and symbols

Description automatically generated

**Model Analysis**

A confusion matrix was used to further evaluate the model's performance. The model achieved high sensitivity for the 'low' and 'high' classes, indicating a high true positive rate for these classes. However, the sensitivity was lower for the 'middle' class.

The heatmap of the confusion matrix shows that the model was better at predicting the 'low' and 'high' price categories, while it had more errors when predicting the 'middle' price category.



**Conclusion**

The SVM model was able to predict the price category of listings with a satisfactory level of accuracy. Future work could involve exploring other machine learning models or feature engineering methods to improve the prediction of the 'middle' price category.

**Data**

**Part 2**

**Data**

**Data**

**Data**

**Data**

**Data**

**Data**

**Data**

**Data**

**Data**

**Data**

**References:**

K, I. P. A. (2019, September 25). Singapore airbnb. Kaggle. https://www.kaggle.com/datasets/jojoker/singapore-airbnb