Case Study

University of Zurich

Department of Banking and Finance

ST 2023

"Optimizing Real Estate Price Class Estimation: An Exploration of Machine Learning Techniques"

Group Members

Victor Nicolae Simtion -victornicolae.simtion@uzh.ch

Lev Maznichenko -lev.maznichenko@uzh.ch

Minh Hien Tran -minhhien.tran@uzh.ch

Anne-Catherine Sophia Walder -anne-catherinesophia.walder@uzh.ch

Summary

Draft!!
Draft of Summary for KNN and Random Forest Classifier Models:
In this study, we have explored the application of machine learning algorithms to predict real estate price classes using a dataset of residential homes in Ames, Iowa. We have built four different models, one implementing K-Nearest Neighbors (KNN) by Victor, one implementing Linear and Quadratic Discriminant Analysis by Minh, another implementing Random Forest Classifier by Anne-Catherine, and one more implementing Support Vector Machines (SVM), which turned out to be the most performant combined with Random Forest Classifier.

1. K-Nearest Neighbors (KNN) Model by Victor:

Victor explored different techniques to improve the system, such as feature scaling and feature selection. He tested MinMaxScaler and RobustScaler, and compared the results with the StandardScaler. The best scaling for the model turned out to be the StandardScaler. Victor experimented with different feature selection methods, including SelectPercentile, SelectKBest, SelectFromModel, and Recursive Feature Elimination (RFE). The highest test score achieved was 0.82 using the SelectKBest feature selection. The trade-off between accuracy and performance was best suited to the Randomized Search classifier. To address potential class imbalance issues, Victor tried stratified sampling for the train-test split and StratifiedKFold cross-validation. However, the results were inconclusive, and the gap between the KNN model score and test results did not consistently reduce.

2. Random Forest Classifier Model by Anne-Catherine:

Anne-Catherine started by encoding categorical variables and dealing with missing values. After splitting the data into training and testing sets, she implemented an instance of the Random Forest classifier. She defined outer and inner cross-validation schemes and a range of hyperparameters for the search space. RandomizedSearchCV was used to perform the grid search and fit the model. The best CV accuracy was found to be 0.83, and the best parameters were as follows:

- n_estimators: 560
- max_features: auto
- max_depth: None
- min_samples_split: 5
- min_samples_leaf: 1
- bootstrap: False

The model's performance was evaluated using accuracy, precision, and recall. The results indicated that the Random Forest Classifier model was effective in predicting real estate price classes.
3. LDA & QDA
Minh started with data pre processing steps such as removing columns with too many values, replacing non-numeric values with NaN and replacing NaNs with zeros. Then he applied feature scaling using StandardScaler, and class imbalance correction using RandomOverSampler.
Next, Minh splitted the resampled data into training and testing sets, performs hyperparameter tuning using GridSearchCV, and selects the best hyperparameters for

LDA and QDA models separately. Then, he fitted the LDA and QDA models with the best hyperparameters, makes predictions on the testing set, and evaluates the performance of the models using accuracy, confusion matrix, precision, recall, F1 score and ROC-AUC (Receiver Operating Characteristic - Area Under Curve).
The accuracy he received for LDA and QDA were 0.8 and 0.829 respectively.

In conclusion, we have developed four models, KNN and Random Forest Classifier, for predicting real estate price classes. The best test score achieved for KNN was 0.82, and for Random Forest Classifier, it was 0.83. Further work is needed to address potential class imbalance issues, and additional algorithms, such as Linear and Quadratic Discriminant Analysis and Support Vector Machines, will be explored in future drafts.