IST 5535: Machine Learning Algorithms and Applications | Deliverable 2:Final Project

Group 2: Tanner Fry, Sayantan Majumdar, Daren Liu

Research Questions and Dataset

In this project, we used the New York rental price dataset. The aim of this project is to find the most significant predictors for house rental prices and the best machine learning model for this dataset. This dataset can be obtained from https://www.kaggle.com/sab30226/zillow-rents-2020. It contains 22 explanatory variables (columns) that detail almost every aspect of Rent prices for apartments in New York listed on the site Zillow.com. There are 6955 observations (rows) in this dataset related to housing rental data in New York including yearBuilt, city, postal code, address, price, area, and more. New York rental dataset is the regression-based dataset. We need to predict the rental price based on the explanatory variables In this project, we are dealing with a multivariate regression problem that is concerned with both prediction and inference.

All methods applied or tried

In our project. multiple linear regression, regression tree model, k-nearest neighborhood method (KNN), random forest model, support vector machine (SVM), and artificial neural network (ANN) were utilized for the regression analysis. Ridge regression could have been used but wasn't as there are many models already in use. For multiple linear regression, univariate Feature Selection, best subset selection, forward stepwise selection, backward stepwise selection, and lasso were applied to find the best linear regression model. Similarly, for the support vector machine, the linear, radial, and polynomial kernels were used to select the best SVM model.

Summary of Methods Compared

Table 1: Performance of the evaluated models for predicting rental prices in New York.

Evaluate Model	Train Set			Test Set		
	RMSE (USD)	R^2	MAE (USD)	RMSE (USD)	R^2	MAE (USD)
Multiple Linear Regression	9360.477	0.971	2143.844	9464.870	0.960	2026.440
KNN Regression	4890.567	0.992	892.930	5253.795	0.987	1086.894
Regression Tree	4891.061	0.992	1373.122	7520.077	0.987	1526.196
Random Forest	4732.344	0.993	1040.033	6969.323	0.987	1144.404
SVM (Radial Kernel)	9091.456	0.981	4129.222	7801.044	0.979	4812.664
ANN 3-layer	4360.192	0.994	1087.822	7808.273	0.971	1385.642
ANN 2-layer	5046.434	0.992	1412.792	6152.277	0.983	1429.834
ANN 1-layer	**4347.197**	**0.994**	**1255.458**	**4948.706**	**0.988**	**1275.619**

As observed from Table 1, the model rankings based on test set performance are as follows: ANN 1-layer, KNN regression, Random Forest, Decision Tree, and SVM. These rankings are taking into account RMSE, R^2, and MAE.

Results

In this section, we briefly discuss the results for ANN 1-layer, KNN regression, random forest, and regression tree. Since ANN 1-layer has the highest performance, we don't explicitly discuss ANN 2-layer and ANN 3-layer in this report. These are explained in the technical appendix along with SVM and linear regression.

From the model comparison, ANN 1-layer has the best performance among all the models. The RMSE and MAE for the ANN 1-layer training model are USD 4347.197 and USD 1255.458. For the ANN 1-layer testing dataset, the RMSE and MAE are USD 4948.706 and USD 1275.619 respectively. Intriguingly, the R^2 for ANN 1-layer is the highest, which is 0.994 for training and 0.988 for testing. The corresponding neural network structure is shown in Figure 1.

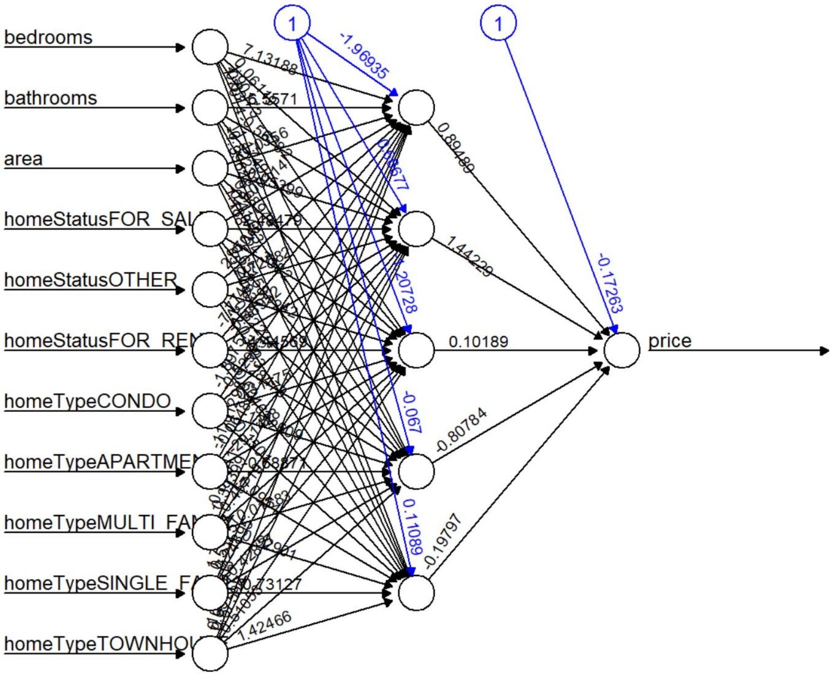

Figure 1: Neural network structure for ANN 1-layer. Both train and test data were scaled. The final predictions were obtained after reverse transforming the predicted price values.

In the KNN regression model. when the number of neighbors is smaller than 20, the increase of RMSE is obvious with increasing the number of neighbors. So the best KNN regression model is k=3, the RMSE of the KNN regression model is USD 4890.567 for training and USD 5253.795 for testing. Moreover, the R^2 for the KNN regression is 0.987. It is noteworthy that we used repeated k-fold cross-validation to get the optimal value of the number of neighbors, k, in the KNN regression model. We observed that both k=1 and k=2 showed better training set performances but these models were heavily overfitting the test data. Additionally, for $k > 3$, the models were heavily underfitting with training R^2 close to zero and test performance metrics better than that of k=3. Therefore, we chose k=3 as the optimal value. Figure 2 shows the cross-validation plot (repeats=3, folds=10) for the KNN regression model.

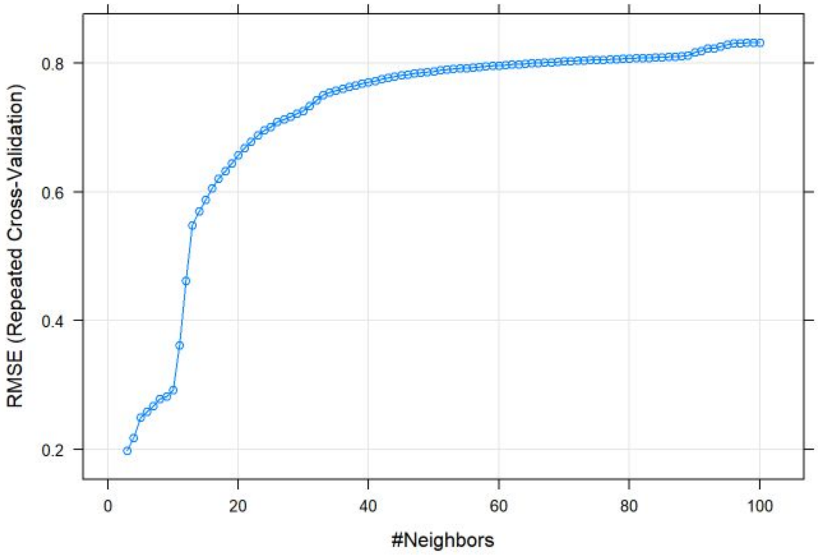

Figure 2: Cross-validation plot showing the change in RMSE with the change in the number of neighbors. Like ANN, both train and test data were scaled and then reverse transformed afterwards.

The next model is the random forest, the graph below in Figure 3 shows the relationship between RMSE and the number of predictors. When mtry>4, the RMSE slows down significantly. The minimum RMSE appears when mtry=10. The minimum RMSE is USD 4732.344 for training and USD 6969.323 for testing. The R^2 for the random forest is 0.993 for training and 0.987 for testing.

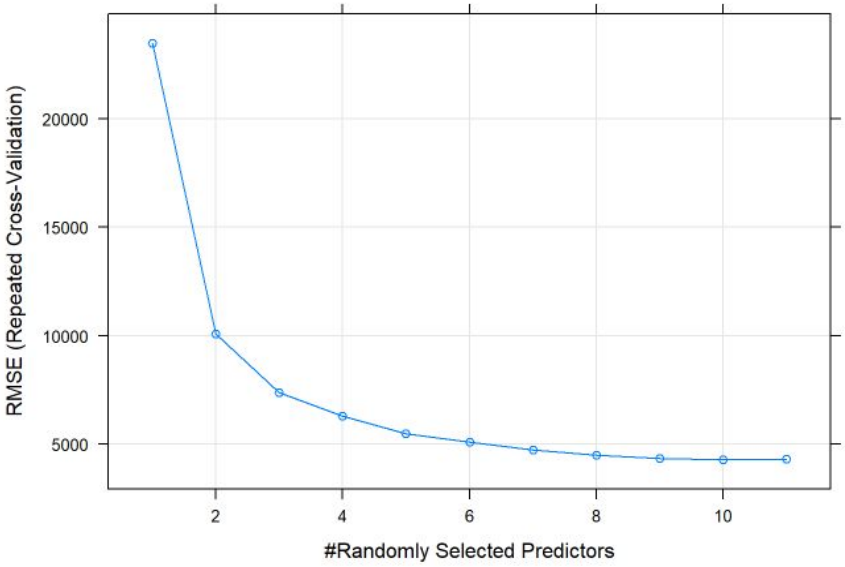

Figure 3: Cross-validation plot for random forest. Here, repeats=3, and folds-10. The RMSEs are in USD.

Another model that is suitable for the dataset is the regression tree which uses the homeStatus and area as root and non-leaf nodes respectively to predict the rental price (terminal node). The RMSE for the regression tree is USD 4891.061 for training and USD 7520.077 for the testing dataset. The R^2 for the regression tree is similar to random forest and KNN regression..

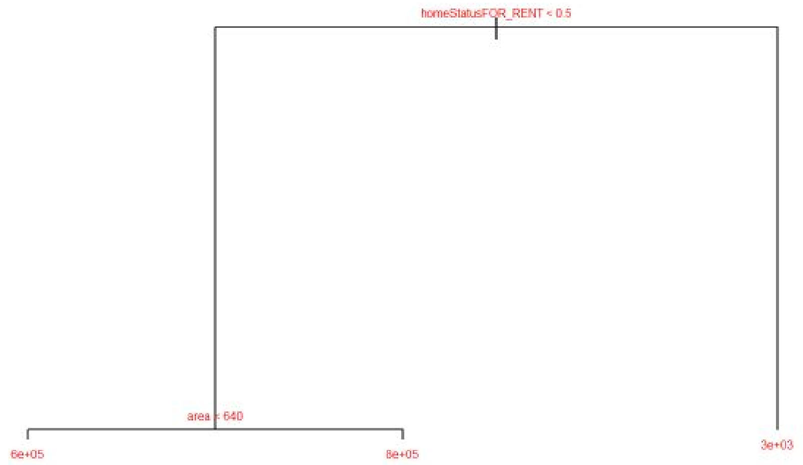

Figure 4: Decision tree for predicting the rental price. Here, pruning was not required.

Conclusion

This project uses the New York rental dataset and demonstrates a practical use-case of both prediction and inference problems in the machine learning domain. Outlier handling is a key component of this project. If the outliers present in this data set are not removed, all the evaluated models summarized in Table 1 failed due to overfitting as there were very few price observations that were greater than USD 1 million or 1e+6 (the variance was also high). So, those observations effectively became noise.

Accordingly, if the threshold is set higher than USD 1e+6, e.g., USD 2e+6, the models are strongly influenced by the extreme values resulting in both train and test scores to be very close to zero. This could be because the expensive properties (having prices more than USD 1 million) are heavily undersampled and effectively act as outliers. Moreover, the median price is around USD 2100. Apparently, log transformation fixes these issues (without any outlier removal), however, when the reverse transformation of the log-scaled price values is performed, the small errors on the log scale become quite large in the original scale.

Nevertheless, we can clarify the important factors that affect the rental price as well as appropriately predict the rental prices using these factors. We found that the ANN 1-layer model has the best performance among all models (Table 1). Since this is a black-box model, we cannot identify the important factors influencing rental prices. Intriguingly, the random forest model shows the feature importances (described in the technical appendix). So the entire approach can be used both for prediction and inference. For example, we found that a rental house has a bigger area and more bathrooms have strong bargaining power for the rental market. Thus, the renter may have less bargaining power if they are looking forward to these factors. Moreover, the rental agent and the owner know which factors can increase the rental price most, and as such can upgrade them accordingly to boost the rental price. Therefore, we can conclude that we have justifiably identified the answers to our proposed research questions and successfully met the desired objectives.