Ryan Curling

Dr. Bryan

BAS 474

12/9/2022

BAS 474- Course Project

Problem statement:

We are seeking to apply data mining algorithms, to understand and predict rental prices from a real-world dataset. Our goal is to determine the most optimal post-graduation city to live in based on difference of income and next year's rent.

Models considered:

This project considered several data mining methods, including generalized linear model, elastic net regression, support vector machine (linear, polynomial, and radial) , Tree model , gradient boosted model, neural net , and knn model. These methods are commonly used for data mining and have been proven to be effective in a wide range of applications. The performance of the different methods was compared using cross-validation, and the best-performing method was selected for further analysis.

Metric used to compare methods that are considered:

The metric used to compare the performance of the different data mining methods in this project is the root mean squared error (RMSE). This metric measures the difference between the predicted values and the true values in the dataset. A lower RMSE indicates that the model is able to make more accurate predictions, while a higher RMSE indicates poorer performance. The more errors a model has, the worse it performs. In this project, the RMSE was calculated using cross-validation, which involved dividing the data into multiple folds and training and evaluating the model on each fold. This helps to reduce the effects of overfitting and provides a more accurate estimate of model performance.

Data description:

The data used for this assignment comes from the file from "Rent_Data_by_Zip.csv". This dataset contains valuable information about median household incomes, rent prices, along with other variables for different zip codes. By analyzing this data, we are able to learn about trends in rent prices in correlation with other predictor variables to predict future rent prices for any zip code within our dataset.

Data Clean-Up and Preprocessing Steps:

Before making visualizations, I needed to determine the predictor variables of nextYearRent. To do this, I fit a model using the linear model function. Upon making this model, I ran a summary of it and determined my predictor variables

Here is the summary output:

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.063e+02  1.499e+01  20.439  < 2e-16 ***
zip_code                    -1.201e-03  2.607e-04  -4.607 4.18e-06 ***
rent_mean                    2.953e-01  4.770e-02   6.191 6.49e-10 ***
rent_max                     6.670e-01  4.787e-02  13.934  < 2e-16 ***
rent_sd                      2.801e-01  8.261e-02   3.391 0.000703 ***
medianHousholdIncomeUSD      4.224e-04  7.967e-05   5.302 1.20e-07 ***
YearMortgageInterestRate309 -5.487e+01  2.859e+00 -19.193  < 2e-16 ***
homePrice_mean              -7.195e-03  7.831e-04  -9.188  < 2e-16 ***
homePrice_max                7.176e-03  7.827e-04   9.168  < 2e-16 ***
homePrice_sd                -9.572e-03  1.233e-03  -7.763 1.01e-14 ***
---
```

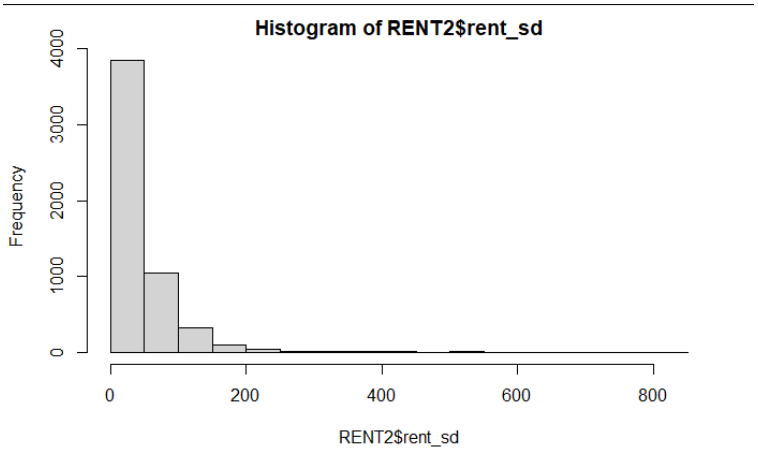
-

My determining factor for these predictor variables was the p-value. As you can see above, all these variables have a statistically significant relationship to nextYearRent. This is what we will fit our predictive model.

<u>Now that our predictor variables have been established, we need to clean the data:</u>

The first step in doing this was checking for skewed columns in our dataset. After running histograms on all columns. I determined; normalization was required.

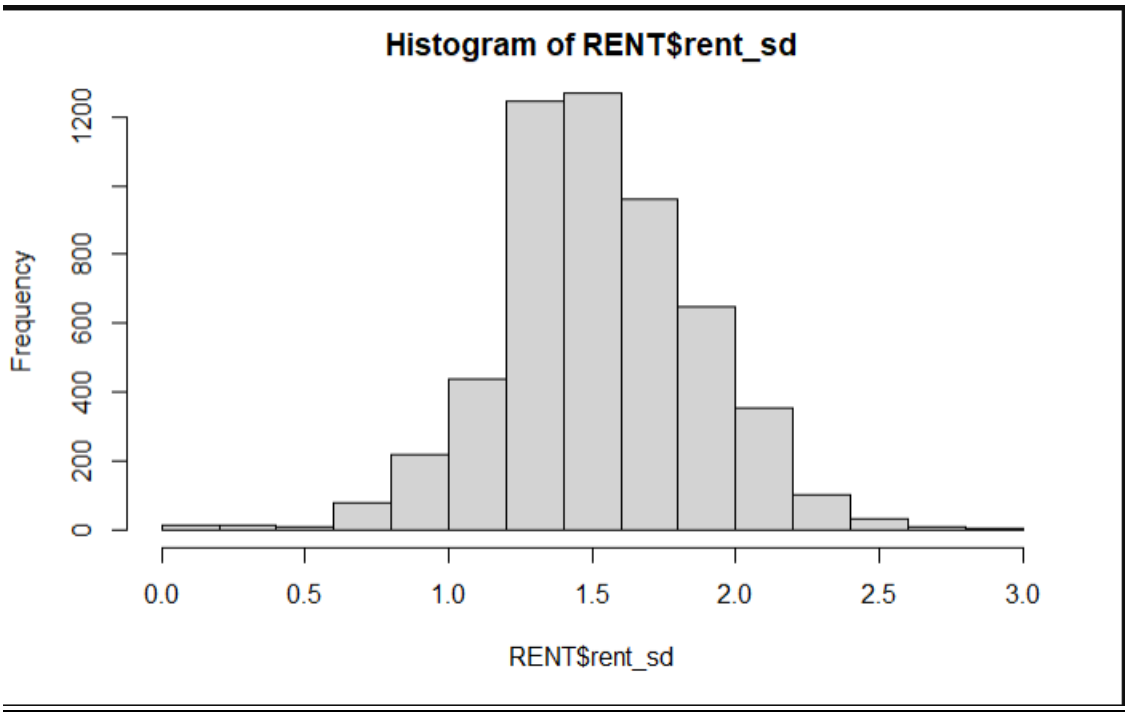
<u>Here is an example of a skewed column 'rent_sd':</u>



Normalization:

After determining the data was partially skewed, it was clear normalization was necessary. By applying log normalization to the data, I transformed the skewed variables into a more symmetrical distribution. For example, taking the log10 of the YearMortgageInterestRate309 variable will compress the high values and stretch the low values, resulting in a more balanced distribution. This can make the data more suitable for modeling and improve the performance of the algorithms.

Here is the same column after normalization:



Removing NA Values:

In addition to log normalization, NA values were removed from our data in order to accurately create our predictive model. A variable called ‘base’ with these new non-NA values was

created. Most machine learning algorithms are designed to work with complete, well-defined datasets, and cannot handle missing values properly. When a missing value is encountered, the algorithm may produce unexpected or incorrect results. Therefore, it was extremely important to identify and remove NA values from the dataset before building a predictive model.

Results:

After modeling all methods, both tree and random forest methods proved to be far more accurate than the other models. The random forest model was found to perform better than the tree model, based on the MAE and RMSE metrics. The random forest model had a lower MAE and RMSE, indicating that it made more accurate predictions than the decision tree model. This suggests that the random forest model is the best choice for this data and can be used as the final method for predicting future rent prices.

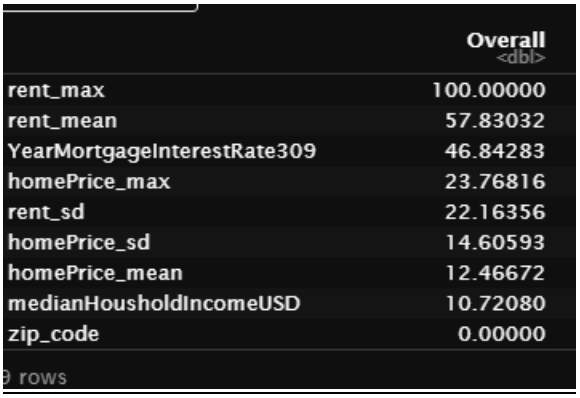
Here is my output for the model I chose:

	mtry <dbl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>
2	4	0.05104157	0.9577529	0.01884004

Interpretation:

Upon fitting a model, I thought it would be appropriate to create a variable importance plot, to see which variables were the most influential to our model.

Output:



	Overall <dbl>
rent_max	100.00000
rent_mean	57.83032
YearMortgageInterestRate309	46.84283
homePrice_max	23.76816
rent_sd	22.16356
homePrice_sd	14.60593
homePrice_mean	12.46672
medianHousholdIncomeUSD	10.72080
zip_code	0.00000

9 rows

Here we can see that rent_max and rent_mean are the most important variables to this model. This makes sense considering we are predicting nextYearRent. Upon making this variable importance plot, I was able to calculate the zip codes with the biggest difference in incomes and next year's rent. This ultimately solved our problem statement:

Top Zip codes:

30327

28207

38139

40059

37027

Executive Summary:

The purpose of this project was to apply data mining algorithms to a real-world dataset in order to predict future rent prices. The dataset consisted of rental and housing data from different zip codes across the United States.

After pre-processing the data, including log normalization and removal of missing values, two models were considered for predicting future rent prices: decision trees and random forests. The performance of the models was compared using mean absolute error (MAE) and root mean squared error (RMSE) as evaluation metrics.

The results of the analysis showed that the random forest model performed better than the decision tree model, with a lower MAE and RMSE. This indicates that the random forest model was more accurate in predicting future rent prices. The top performing zip codes, as identified by the model, were 30327, 28207, 38139, 40059, and 37027.

Overall, this project demonstrates the potential of data mining algorithms for predicting future rent prices. The use of the random forest model was shown to be effective in identifying the top performing zip codes and making accurate predictions. These results can be useful for informing future decisions and strategies related to rental properties.

Appendix:

https://stackoverflow.com/questions/tagged/r

https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/

https://www.geeksforgeeks.org/data-cleaning-in-r/