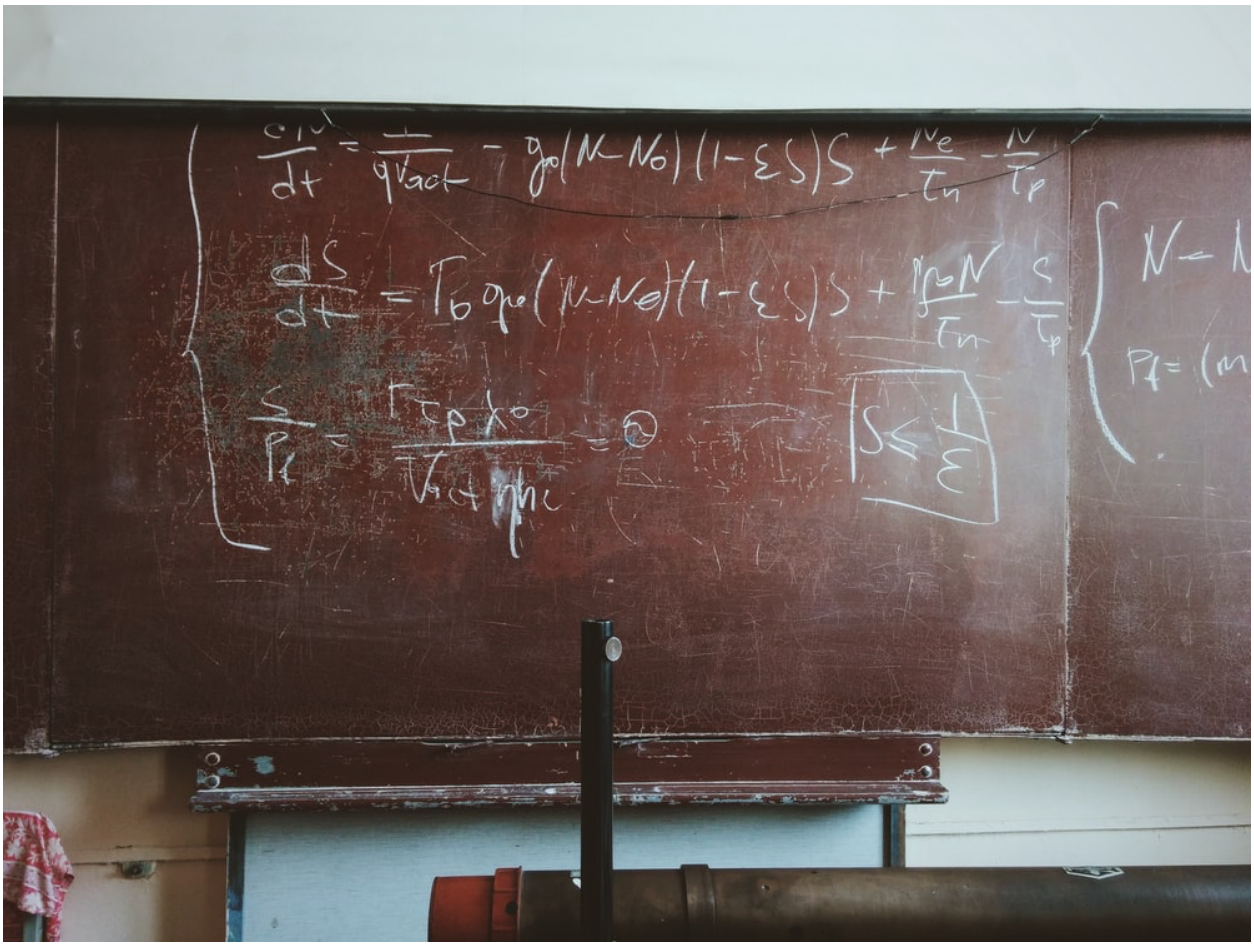

Take-Home Data Science Project with Chi-Squared and Machine Learning Models

Devin Powers

March 30th , 2021

Abstract

In this paper, I will attempt to answer the questions given in the Home Partners of America Data Assessment. The dataset given in the assessment shows whether an existing lease on a home was renewed or not renewed. By training the data on multiple models, I found the best model to be either the decision tree classifier or logistic regression. I found that the biggest impact on whether someone renews a lease to be 1. no change in rent and 2. no fines or violation. Both of those features correlated on the Heatmap, Chi-Squared Values, and in my Machine Learning Algorithms in part 2 of the assignment.

1 Introduction

From the data, nearly 20% of the customers resigned their lease as shown in **Figure 1** below.

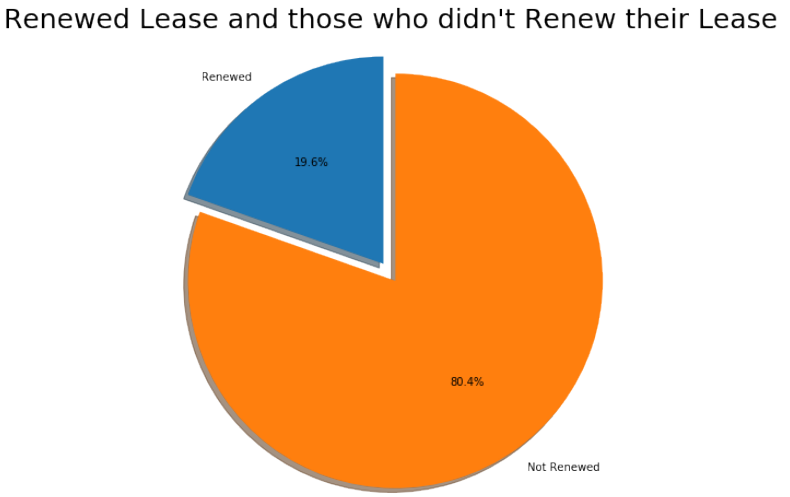

Figure 1: Pie Chart of Renewed Leases

2 Pre-Processing the Data

Table 1: Columns and Descriptions of the Dataset

Column	**Description**
lease_id	Unique ID for each lease
no_rent_change	The lease rent was not changed for the next term
rent_change_10	The lease rent was changed by at most 10% for the next term
rent_chnage_20	The lease rent was changed by more than 10% but at most 20%
lease_length_1	The resident has lived for at most 1 year on the lease
lease_length_2	The resident has lived for more than 1 but at most 2 years on the lease
lease_length_3	The resident has lived for more than 2 but at most 3 years on the lease
age_range	Set of columns indicating average age of residents in the household
NoFinesViolations	Indicates the resident had no fines or violations on the lease
PositiveSurvey	Indicates the resident provided a positive feedback on the survey
LatePayments	Indicates the resident has never been late on making payments
HOA_mandatory	Indicates whether there's a mandatory HOA fee on the lease
Renewed	Indicates whether the resident renewed the lease or not

From the dataset given, the only irrelevant column to drop from the data analysis is the lease_id. The dataset was given in binary (0 or 1), there were no NULL values in the dataset.

3 Chi -Square Test for Independence

Chi-Squared Test for Independence was picked for the data analysis because it allows for comparing two categorical variables and test if they're related. In the case of the dataset, the Chi-Square Test for Independent allowed for comparing every feature (column) with the renewal column.

The first step for performing the Chi-Squared for Feature Selection was stating the Hypothesis.

- **Ho**: There is no relationship between the two categorical variables (with one being Renewed). (They are independent.)
- **Ha**: There is a relationship between the two categorical variables (with one being Renewed). (They are not independent.)

The next step was building the contingency tables using crosstab function in Pandas. The Contingency table is shown in **Figure 2**, and the corresponding matching Heatmap is shown in **Figure 3**.

Renewed	0	1	All
no_rent_change			
0	51797	10372	62169
1	12411	5270	17681
All	64208	15642	79850

Figure 2: Contingency Table for No Rent Change Column

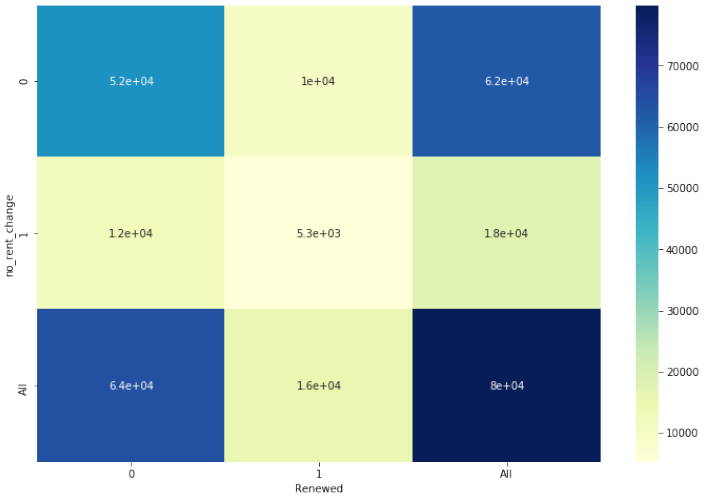

Figure 3: Heatmap for No Rent Change Column

Once the contingency tables were constructed for each column, they were passed into the function categorical_dependency and the function returns the Chi-Square Statistic, P-Value, and if whether the column rejected the Ho (dependent) or fail to reject the Ho (independent).

The only column that failed to reject the H0 was the column with *age_range_24_29*. Since the column failed to reject the H0 it was, I discarded the feature value from building the model in part 2 of the assignment. In **Figure 4**, a plot of the highest Chi-Square values is shown.

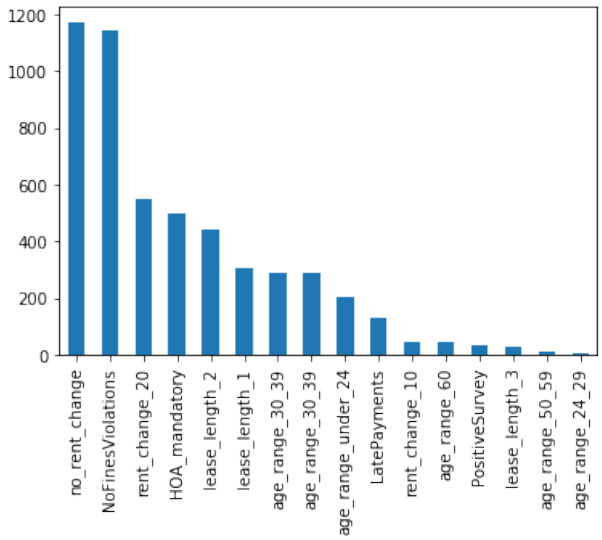

Figure 4: Graph of Chi-Square Values

As we can see above, the higher the Chi-Square Value, the higher dependence it had on the Renewed column.

After completing the Chi-Square Test for Independence, I performed a simple correlation on the dataset and a Heatmap is shown below in **Figure 5**. A table of the highest correlation (> .08) from the Heatmap is shown in **Table 2** below.

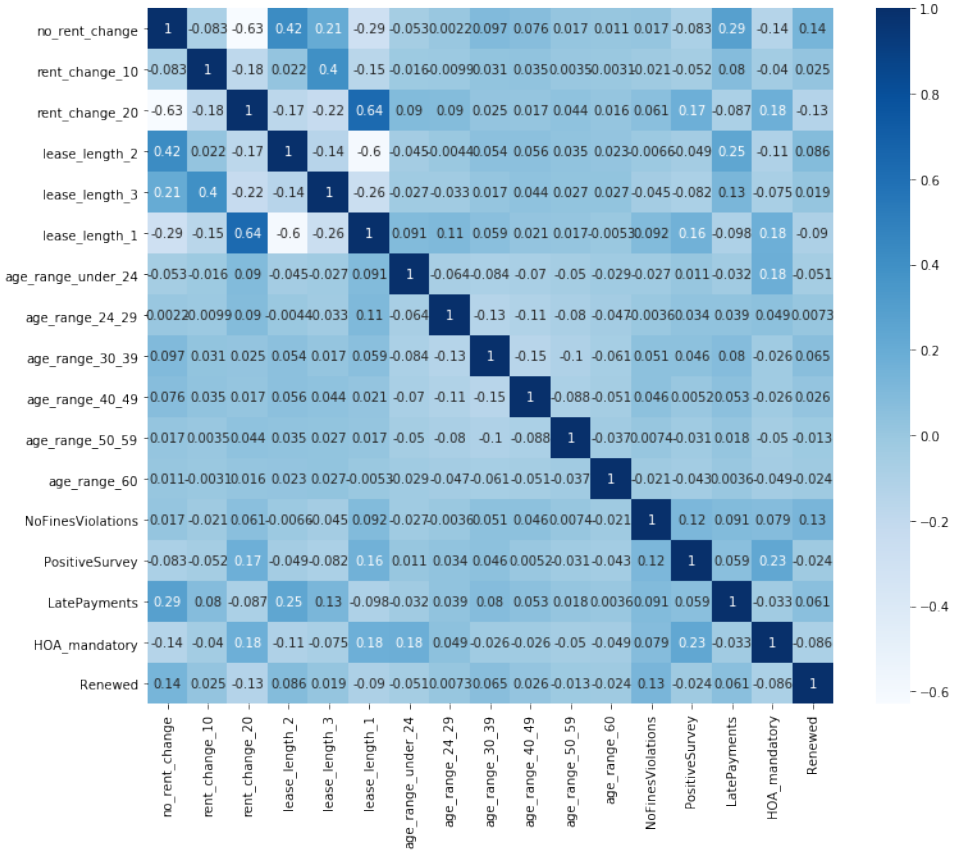

Figure 5: Heat Map of Correlation

Table 2: Relevant Features from Correlation

Feature	Correlation Value
no_rent_change	0.137282
NoFinesViolations	0.128865
rent_change_20	0.127915
lease_length_1	0.089716
HOA_mandatory	0.086500

Part 2

4 Machine Learning Models

For Part 2, building the model, I used 3 different models from the sci-kit learn library and calculated their accuracy scores. Before splitting and training the data, I dropped both the *lease_id* and *age_range_24_29* from the dataset. I split the data into 70% train and 30% test. Note that feature selection is the most important part of building machine learning models and irrelevant or partially relevant features can negatively impact model performance.

4.1 K-Nearest Neighbors

The first algorithm used on the data was K-Nearest Neighbors, the KNN algorithm assumes that similar things exist in close proximity. After completing the algorithm, I tested the accuracy of the model, which is just the fraction of samples predicted correctly. The accuracy score was found to be 0.68227.

4.2 Decision Tree Classifier

The second algorithm used on the data was the Decision Tree Classifier and the accuracy score came out to be 0.80409. After completing the tree, I used the **feature_importances_** tool to help find the important coefficients that helped the tree build the model. A figure of the Feature Importance is shown below in **Figure 6**. The most important features included no*_rent_chan*ge and *NoFinesViolation*.

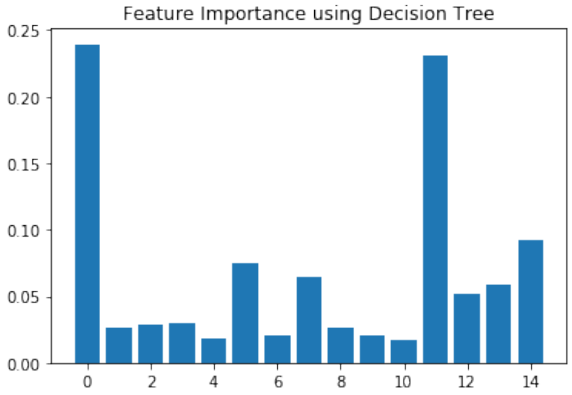

Figure 6: Important Features

4.3 Logistic Regression

The last algorithm I used the data was Logistic Regression and the accuracy score was found to be 0.804717, which is very similar to the accuracy score using the Decision Tree.

5 Summary and Conclusions

After performing both the Chi-Square Test for Independence, Correlation, and different Machine Learning algorithms on the dataset I would recommend to Home Partners of America that the most important feature in deciding if someone renews their lease is is 1. No Change in Rent and 2. No Fines or Violations. For the detractors for whether someone renews a lease is age ranges 24-29 and 50-59. Intuitively they both make sense, someone in the age range 24-29 are young in their careers and might be actively moving cities and in the age range 50-59, they might be downsizing from their current place.