Daniel Mejia

11/3/2021

Milestone 5 – REPORT

Our problem started with crime and how home prices around the area affect the possibility high crime. The problem was straight forward and required data analysis at the core to answer the possible question. Libraries used for the analysis were extensive and did make the difference in answering the question and obtaining the solution. However, the solution is not always what is expected. For the project below the problem was simple, how can prices home around an area determine crime rate. The analysis can be used to solve and answer many problems in the home investment, crime, and possible deterrence used by the police enforcement.

The data chosen for the project below involved the Boston House Data. Just as the name implies the data contained information about Boston included in the report was crime rate, home prices, and many other interesting features. One of the most useful features were the Ages and taxes because of the correlation with people financial status. Using the Boston data, finding the correlation for crime and money was the goal. But finding correlation and visual parity with other features was also required in retrospective.

The analysis started with Age and Crime, the visual 2D was complex and impossible to find any correlation. The lines ran all over the X and Y coordinates. No pattern was even remotely visual from the graph. My beginning was horrendous and let to nowhere. I did not know how to even graph the data with so many points and data running all over the place. The result was the graph below with no answers, actually more question was concluded from the graph.

Chart, line chart

Description automatically generated

I moved to visualize Age and Price to be used to find the correlation in Crime. The results were better and screamed a correct pathway to continue. I continued with the scatter graph technique to visualize correlation between data. Except for the outliers in at the edges of the graph, the data seems very plausible for use in the detection of Crime.

Chart, scatter chart

Description automatically generated

I moved into the next part correlation and getting data with valuable meaning from the Boston data. Also, the following analysis was going to help me remove features based on the correlation with the other data. Before using any automatic steps, manual steps are taken as well to make the model better. The correlation led me to decide to remove some features like the CHAS, B, PTRATIO as a test to improve the model. However, removing features manually prove unfertile and did not affect the model.

Chart, bar chart

Description automatically generated

Feature selection was done manually and logically in the code. The picture below shows the steps taken to complete the removing of features and selecting data to be process for the model. Removing the not useful features made the model faster and less intensive on the memory used. But the truth purpose of finding a solution sending my correlation to above 95% was not found by removing features.

Text

Description automatically generated with medium confidence

For the model selection, the RandomForestRegressor was chosen. I decided on the model because of the complexity of the data and knowing the previous model was only outputting 50% correlation. I was required to find a better model to solve the problem faced by me. I found the model manually reading different models and deducing the best outcome based on the description. Trial and error happened some models which I believe were going to work did not at all.

My conclusion was conclusive, using the data for prices, ages, and other features, we were able to determine crime at a 93%. Most people would say that money runs the world and indeed it runs mostly how people move and what people are around certain areas. Businesses would care about the data because determining crime and being able to see the outcome base on age and salary is important to purchase homes. Also, determining how a neighborhood might change based on pattern of changing in home prices and salary is also a game changer. The data can be used in different ways to turn houses or knowing when to get out of homes around an area if income and home prices change drastically. Also, the police can use the data to build station close to certain areas changing from low crime to high. The code below shows the model and the graph correlation.

Chart, scatter chart

Description automatically generated