Forecasting Future Walmart Sales using KNN Regression

Muhammad Salman – 100995527 - Masters of Engineering in Electrical and Computer Engineering

Seyedeh Marziyeh Zamiri – 101214196 - Masters of Engineering in Electrical and Computer Engineering

Professor Mikhail Genkin, DATA 5000 Y

March 11th, 2021

Introduction:

Using a provided Walmart dataset from Kaggle, data engineers can predict future sales that the megastore will generate. The information in these datasets comprise of a wide range of information regarding 45 stores and their localities. These include departments per store, weekly sales, sales during weeks of major holidays, physical size of the store, unemployment rate for the local population, etc. This is quite significant as it would allow the store to predict which departments within certain stores would yield greater sales depending on the locality and time of year. In a realistic sense, this would allow Walmart to calculate how much and what kind of inventory they would require, leading to a growth of sales and minimizing of costs.

Methodology/Discussion:

The dataset comprised of CSV spreadsheets which are easily accessible in Excel. All data analysis for this project past this point was done using Python3.9. Within Python, the Pandas library was used to attain quick access to CSV files and build a virtual data frame using the given the data. The various data frames from each CSV file were first cleaned, as is done in every data science project. In simple terms, data that is N/A can be filled with 0 or any other value using Pandas' commands. Once that was complete, the CSV files were merged into one large data frame for combined access.

The next stage involves exploring the data, and observing how certain features relate to others. For each comparison, a certain type of plot was generate using the Matplotlib and Seaborn libraries. An example of this was a bar graph that was generated to see the difference in sales between holiday weekly sales and standard weeks.

Results:

The final stages will include modelling the data, training the model, and verifying the model. Although this is yet to be performed, a regression model will be used, since the predictable value is a dependent variable, weekly sales, based on a series of independent variable, such as store size, holidays, locality statistics, etc. To be more precise, this model will be based off the K Nearest Neighbors algorithm, as it will allow us to also classify data points, then characterize them. This algorithm trains based off similarities between existing features.

Conclusion:

If we are able to generate a successful sales forecasting model, it will be significant for the retail industry, and even the wider consumer market, as business owners will be able to project future revenues much in advance, leading to higher organization and profits.