INFO 659 Introduction to Data Analytics

Group Project

Group Number **8**

Team Members
1. Anusha Narayanan
2. Manisha Nandawadekar
3. Siva Prasath C S
4. Amit Nijsure

Abstract

Businesses are often at crossroads with crucial decisions that need to be implemented rapidly and the outcomes of these decisions will determine whether a business will live to see another day or will resurface from tremendous losses and fly high again with profits. Businesses always run the risk of unnecessary stock pile up or keeping stocks at low level which ultimately has an adverse effect on the revenue. Producers and shopkeepers face tremendous difficulties in determining the factors that affect the market sales and end up suffering losses due to bad decision making. Some of these factors that are proven to affect sales are unexpected weather conditions, holidays, commodity price inflation, natural calamities, and less footfalls. For example, an owner of a car repair depot would stock up on car parts expecting a rise in customer service requests after the Christmas Holidays but the hurricane forces a people to stay at home thus, affecting the owner's overall business. It is necessary to come up with a prediction model to overcome this problem, which can take these factors into account and provide sales prediction given a certain set of factors.

In this project, we will be using Kaggle's Walmart Store Sales Data and try to find a correlation between a potential factor that could affect sales and actual sales. These correlations will then be merged to form a decision making or a supervised machine learning model that predicts the amount of sales based on certain set of factor values. Thus, the *Weekly Sales* will be the numeric variable that needs to be predicted based on *predictor variables* such as weather conditions, fuel prices, consumer price index and unemployment levels. Following is a detailed description of the dataset –

1.	Dataset	Retail Analysis with Walmart Store Sales Data	
2.	Number of Records	6,436	
3.	Attributes	Store	Integer (Numeric) – Might Not Be Used
		Date	Date
		Weekly Sales	Float (Numeric)
		Holiday	Boolean – To denote a holiday week
		Temperature	Float (Numeric)
		Fuel Price	Float (Numeric)
		Consumer Price Index (CPI)	Float (Numeric) – Average change in the prices paid by consumers for a market basket
		Unemployment	Float (Numeric) – Higher unemployment should ideally lead to lower sales

All the attributes described above will be a part of our analysis task except for *Store* which does not provide a meaningful relationship with the weekly sales or any other attributes. Our task will be to analyze the correlation between each attribute with the target variable *Weekly Sales*. Based on these analyses, we will use a Regression Model, Naïve Bayes Classifier, and a Decision Tree generate predictions and then evaluate model performances by fine tuning the model parameters. The dataset will be split into train and test sets of in the ratio of 80% to 20% respectively and *k-fold cross validation* will be used to obtain better representative results.