DATA SCIENCE EAST AFRICA

DATA SCIENCE PEER PROJECT

Day 11 /20

1). Walmart Sales Forecasting.

Data Science plays a huge role in forecasting sales and risks in the retail sector. Majority of the leading retail stores implement Data Science to keep a track of their customer needs and make better business decisions. Walmart is one such retailer.

Problem Statement: You are supposed to analyze the Walmart Sales Data set in order to predict department-wise sales for each of their stores.

Data Set Description: The data set used for this project contains historical training data, which covers sales details from 2010-02-05 to 2012-11-01. For the analysis of this problem, the following predictor variables are used:

1. Store – the store number
2. Dept – the department number
3. Date – the week
4. CPI – the consumer price index
5. Weekly_Sales – sales for the given department in the given store
6. IsHoliday – whether the week is a special holiday week

By studying the dependency of these predictor variables on the response variable, you can predict or forecast sales for the upcoming months.

Logic:

1. **Import the Data Set**: The data set needed for this project can be downloaded from https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data.

2. **Data Cleaning**: In this stage, you must make sure to get rid of all inconsistencies, such as missing values and any redundant variables.

3. **Data Exploration**: At this stage, you can plot boxplots and qplots to understand the significance of each predictor variables.

4. **Data Modelling**: For this particular problem statement, since the outcome is a continuous variable (Number of sales), it is reasonable to build a Regression model. The Linear Regression algorithm can be used to solve such problems since it is specifically used to predict continuous dependent variables.

5. **Validate the model**: At this stage, you should evaluate the efficiency of the data model by using the testing data set and finally calculate the accuracy of the model by using a confusion matrix.

2). Chicago Crime Analysis.

With the increase in the number of crimes taking place in Chicago, law enforcement agencies are trying their best to understand the reason behind such actions. Analyses like these can not only help understand the reasons behind these crimes, but they can also prevent further crimes.

Problem Statement: You are supposed to analyze and explore the Chicago Crime data set to understand trends and patterns that will help predict any future occurrences of such felonies.

Data Set Description: The dataset used for this project consists of every reported instance of a crime in the city of Chicago from 01/01/2014 to 10/24/2016.

For this analysis, the data set contains many predictor variables such as:

1. ID – Identifier of the record
2. Case Number – The Chicago Police Chain RD number
3. Date – Date of the incident
4. Description – Secondary description of the IUCR code
5. Location – Location of the occurred incident

Logic:

1. **Import the Data Set**: The data set needed for this project can be downloaded from https://www.kaggle.com/currie32/crimes-in-chicago.

2. **Data Cleaning**: In this stage, you must make sure to get rid of all inconsistencies, such as missing values and any redundant variables.

3. **Data Exploration**: You can begin this stage by translating the occurrence of crimes into plots on a geographical map of the city. Graphically studying each predictor variable will help you understand which variables are essential for building the model.

4. **Data Modeling**: For this particular problem statement, since the nature of crimes varies, it is reasonable to build a clustering model. K–means is the most suitable algorithm for this analysis since it is easy to build clusters using k–means.

5.**Analyzing patterns**: Since this problem statement requires you to draw patterns and insights about the crimes, this step mainly involves creating reports and drawing conclusions from the data model.

6. **Validate the model**: At this stage, you should evaluate the efficiency of the data model by using the testing data set and finally calculate the accuracy of the model by using a confusion matrix.

3). Text Mining.

Having a Text Mining project in your resume will definitely increase your chances of getting hired as a Data Scientist. It involves advanced analytics and data mining that will make you a skilled Data Scientist. A popular application of text mining is sentiment analysis, which is extremely useful in social media monitoring because it helps to gain an overview of the wider public opinion on certain topics.

Problem Statement: You are supposed to perform pre–processing, text analysis, text mining and visualization on a set of documents using Natural Language Processing techniques.

Data Set Description: This data set contains scripts of the famous Star Wars Series from the Original Trilogy Episodes i.e., IV, V and VI.

Logic:

1. **Import the data set**: For this project, you can find the Data set on https://www.kaggle.com/xvivancos/analyzing-star-wars-movie-scripts/data.

2.**Pre-processing**: At this stage in a text mining process, you must get rid of inconsistencies such as, stop words, punctuations, whitespaces, etc. Processes such as lemmatization and data stemming can also be performed for better analysis.

3. **Build a Document-Term Matrix** (**DTM**): This step involves the creation of a Document-Term Matrix (DTM). It is a matrix that lists the frequency of words in a document. On this matrix, text analysis is performed.

4. **Text Analysis**: Text analysis involves analyzing word frequency for each word in the document and finding correlations between words in order to draw conclusions.

5.**Text Visualisation**: Using histograms and word clouds to represent significant words is one of the important steps in text mining because it helps you understand the most essential words in the document.

Simple guide to confusion matrix :

- https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
- https://www.youtube.com/watch?v=8Oog7TXHvFY
- https://intellipaat.com/blog/confusion-matrix-python/

Data Science end to end project (California housing price predictions)

- https://www.youtube.com/watch?v=kUsNb_gOo_s

All the best,

Regards Data Science East Africa.