Google Analytics Customer Revenue Prediction

Predict how much GStore customers will spend

Solution implementation

This report details the steps that we undertook in implementing the data preprocessing and attribute analysis stages of the project. It also contains an overview of the course of action for the remaining part of the solution for the project.

Data Preprocessing

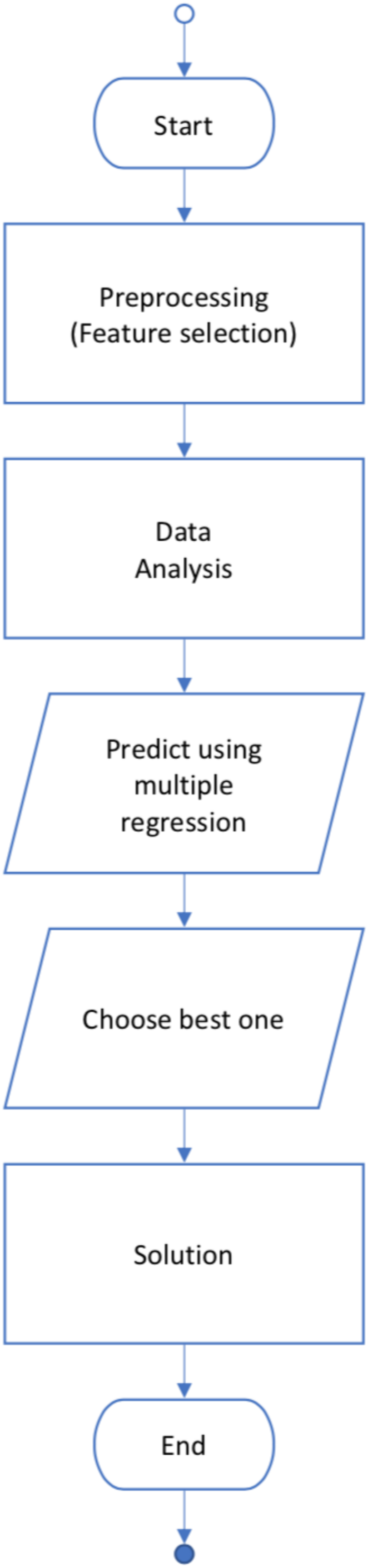
The dataset had some attributes which had subcolumns. We implemented the python code that extracted the subcolumns into individual attributes and the resulting dataset was a data frame named train_data with single column values for each instance. We also implemented the code to format the date attribute. We decided to drop rows that had missing values by checking the values for transactionRevenue attribute, and since there were only a few, dropping them would not have a significant bearing on the resulting model. The 'socialEngagementType' attribute had only 1 value (variance of 0); we therefore automatically dropped it without the need to perform any further analysis.

Attribute Analysis

We performed an analysis of a few selected attributes to determine their importance in predicting the target value. The motivation for the analysis is to give insight into the general underlying structure of the data and also into the relevancy of those attributes in predicting the target value. The attributes are channelGrouping, devices, browser, deviceCategory, operatingSystem, continent, visitNumber, transactionRevenue and source(the source of trafficsource). We have uploaded diagrams that contain the results of each of the selected attributes. These diagrams help us to make some hypothesis like when there is increase in number of visits, the revenue increases in comparison to the neighboring revenue.

Future Tasks

In the next step, we are going to apply multiple regression algorithms on the preprocessed dataset. We will use linear regression and its regularization versions, Decision Tree Regressor and some ensemble algorithms such as the Random Forest Regressor. We will analyze the performance of each algorithm and then select the model that has the best performance. We would then use a threshold value p, for the target value such that the instances with predicted target values > p, would be enlisted in the marketing target group. The flow diagram below shows the steps that we will undertake to complete the remaining part of the solution.


Start
Preprocessing (Feature selection)
Data Analysis
Predict using multiple regression
Choose best one
Solution
End