**Data Scientist Candidate Test**

**Objects:**

This project is to test for the basic knowledge and skills of a data scientist. It consists of two sub projects: one requires the building of machine learning models to classify text files, the other requires one to visualize and analyze historical data to predict spend for next 5 years.

**Data:**

Parts Spend file is the historical Parts spend since 2016. Table 1 explains the column names in the file.

|  |  |
| --- | --- |
| **Column Name** | **Explanation** |
| ORG | Different plant ID |
| Transaction.Date | The date that the parts are consumed |
| Item | Parts SKU |
| Description | Descriptions for parts SKU |
| Quantity | Amount of consumed |
| Transaction.Value | Dollar amount of consumed |
| Group | Category of the Item |
| Line | Manufactory Line that consume the Item |
| Start.Date | The date that the Manufactory Line started |
| Line.Age | How old is the Line when it consumed the Item |

**Project 1:**

In the Parts Spend file, you will notice that some Items are missing information in column Group. The information in column Group was labelled based on column Description. The items without group information belong to GROUP.A, GROUP.B, GROUP.C, GROUP.D, GROUP.E, GROUP.F, GROUP.G, GROUP.H, GROUP.I, and GROUP.J. You need to build a machine learning model to classify these items into different groups.

**Project 2:**

The Parts Spend file includes details of consumed parts for specific lines and the line ages. Niagara is planning to start a new manufacturing Line, and the CapEX director would like an annual budget of parts spend for the next 5 years. Would you please show the director the historical pattern of the parts spend and predict the next 5 years parts spend based on the data?

**Requirements for Senior Data Scientists:**

You can use R or Python to implement the tasks. Send the codes and the following comments back to HR.

For Project 1, please write comments about why chose the models, how to train and evaluate the models, and how to improve the models.

The grouped results are in excel doc “result-project1” under “PredictedGroup” column.

**Why Naïve Bayes**

Naïve Bayes and Support Vector Machines (SVM) are the most common machine learning methods for text classification. Concerning Naive Bayes has low computational cost and is good for handling multiple classes, I chose to use Naive Bayes here.

**How to clean data**

-Replace all the NAs in the Group Column with “ “

-Extract all grouped rows as a new dataset “grouped”

**How to choose train and test data**

-Divide grouped data into training set(80%) and test set(20%)

**How to train and evaluate models**

-Vectorize “Description” column of the training set to get a dictionary of words

- Vectorize “Description” column of the test set as well as the complete dataset to get their dictionary of words

-Train the model on the training set using vectorized description and group

-Use above model to predict group for the test set

-Use confusion matrix and classification report to see model performance

**How to improve model performance**

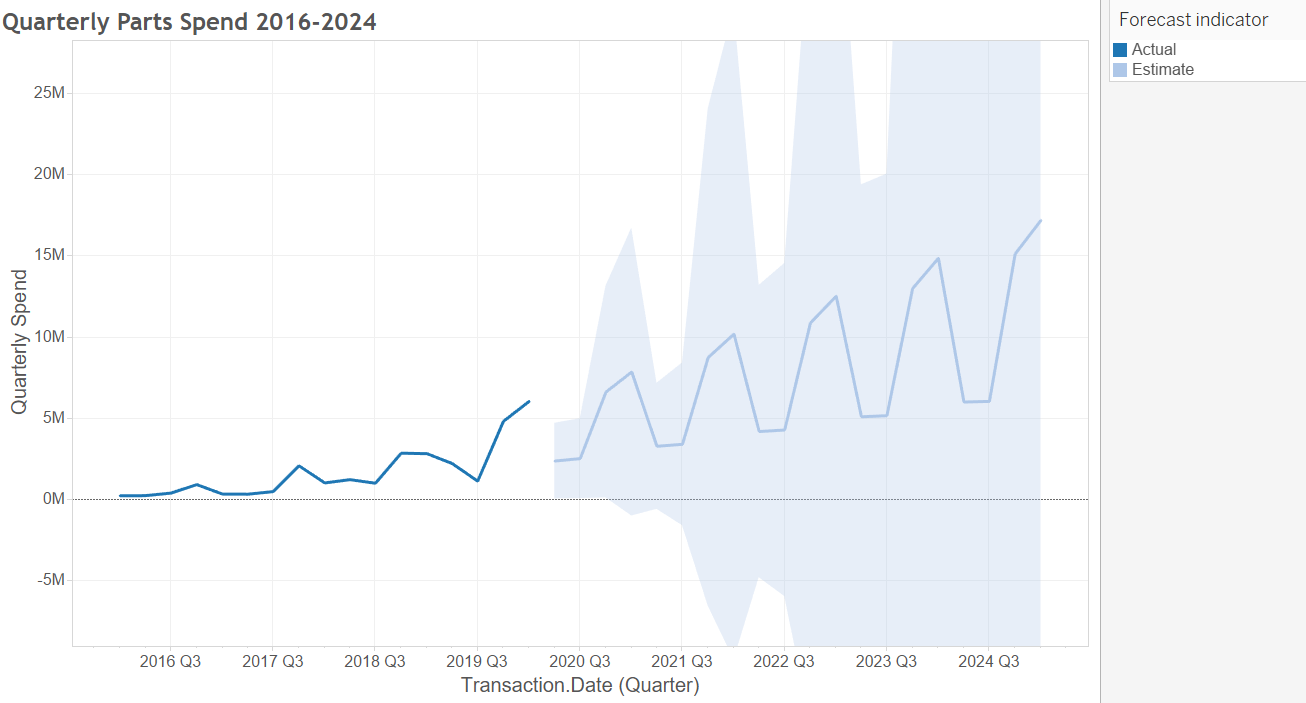
Since the accuracy rate of my model is 99.8%, I don’t think there is need to improve my model here.

But usually when we need to improve model performance, we can

1. Tune classifier’s parameter
2. Apply some classifier combination technique such as boosting, bagging etc
3. Refine the features selected from the data or improve basic parsing

For Project 2, please provide comments and plots of the patterns and tables of the next 5 years parts spend budget.

Below graph shows that from 2016-2019, parts spend usually peaked in quarter 4 and then dropped in the coming quarter 1 of next year. However, in quarter 1, 2020, parts spend didn’t follow the pattern as usual. Instead, it showed an obvious increase.



Based on the total parts spend by quarter from 2016 Q1 to 2020 Q1, Tableau predicts that future quarterly spend will probably follow the pattern as the light blue line. And the detailed yearly total spend accordingly is estimated as below table:

|  |  |
| --- | --- |
| Year | Total Spend |
| 2020 | 17,576,244 |
| 2021 | 23,295,594 |
| 2022 | 29,541,372 |
| 2023 | 35,787,150 |
| 2024 | 42,032,927 |

Please feel free to write comments on any other issues related to the projects.

**Requirements for Junior Data Scientists:**

You can use R or Python to implement the tasks. Send the codes and the following comments back to HR.

For Project 1, you will build a supervised text classification machine learning model. The column Description is input, and the column Group is output. Based on the machine learning model, explain how to clean the data, how to choose training and testing data, how to use a confusion matrix to evaluate the model performance, and how to identify model overfitting issues based on the model performance evaluation.

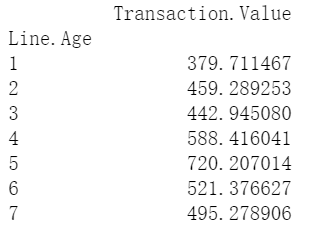
**How to identify model overfitting issues based on the model performance evaluation**

Generalizability suffers in an overfitting model. Therefore, by running the model on the test dataset (20% of the original data), I can detect overfitting by seeing whether my model fits the test dataset as well as it fits the training set. If it doesn’t fit the test dataset, then there exists overfitting.

For Project 2, you need to calculate average parts spend for each line age, then visualize specific line spend vs. average parts spend to remove outliers. After you clean the data, you need to explore the relationship between average parts spend and line age, and predict 5 years of parts spend. Please provide comments and plots of the patterns and tables of the next 5 years parts spend budget.

Please feel free to write comments on any other issues related to the projects.

**Average part spend by line age**



**Relationship between average parts spend and line age**

Based on the result of my linear regression model, as line age increases by 1 unit, spend will on average increase by about $74.7.