Performance Diagnosis Project

In this performance Diagnosis Project, I created an algorithm that could automatically perform a business performance analysis. The first step of data analysis was data preprocessing. I first received raw data obtained from our client stores POS system and preprocessed the data so that it could be analyzed effectively. The raw data comprised of the specific name of the branch, order date, order time, refund time, product category, product name, product price, and order numbers for each order. The first problem was that the order data was blank except the first row so in order to categorize sales by date, I used the fillna() method to fill in the top date to the corresponding sales so that I could group sales by date.

Moreover, I dropped rows that are unneeded by using the .drop method and also removed items that are unnecessary for data analysis such as birthday candles, Ice, coffee for employees, etc. I also replaced items that are redundant such as Americano(brewed) and Americano. After preprocessing raw data, I conducted time series analysis such as Arima and Sarima in order to predict future sales and I also used statistical tools such as linear regression and random forest to predict future sales. In order to enhance the accuracy of the time series analysis, I calculated the MSE(marginal squared error) and MAE(marginal absolute error) between the predicted value and the actual value.

Also, I have developed an Apriori algorithm that could predict future sales effectively. The Apriori algorithm is a tool that determines the top item combinations that could be sold based on sales history. When using the Apriori algorithm, three metrics related to association rules which are support, Confidence, and lift. Support is a measure that gives an idea of how frequent the itemset is in all the transactions and is calculated by the following formula.

Support is the probability of a certain combination to occur out of the entire number of transactions as illustrated on the formula below.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

Confidence is a measure that defines the likeliness of occurrence of the consequent given that the combination already has antecedents.

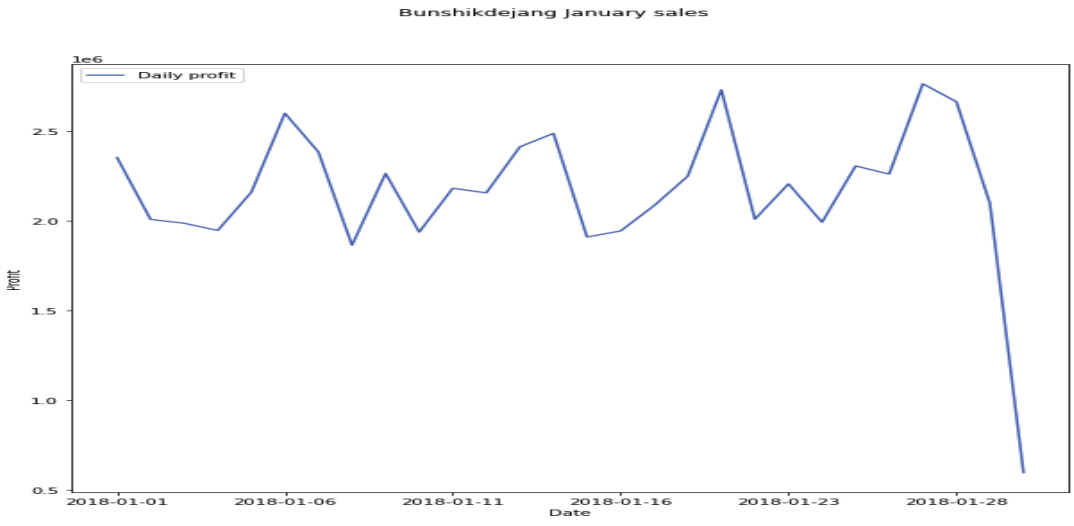
$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

Lift is a measure that controls the support of consequent while calculating the conditional probability of occurrence of {Y} given {x} and is calculated by the following formula.

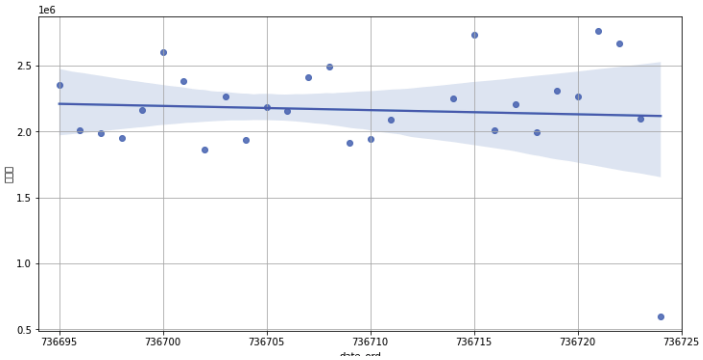
$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

An apriori algorithm is an algorithm that filters all values that are above a certain support, confidence, and lift value and it usually shows the top 10 combinations that are most likely to occur. Based on the menu combination options selected, stores can decide easily on which menu combination they could sell and give promotions.

Moreover, I have also made several visualizations that could represent data. The image below shows the January sales of a snack store called Bunshik Leader. The x axis represents the date and the y axis represents the sales in Korean Won.



The image below is a graph that shows the correlation between sales and date. This graph was created through a lambda function.



I have also used the Pearson`s Correlation in order to calculate the sales trend for a specific branch. If a Pearson`s correlation is negative, it indicates that sales are declining and if the Pearson`s correlation is positive and over 0.5 it means the sales are increasing as time goes by which is a positive indicator of the business` performance. The code below uses pearson`s correlation to calculate the correlation between date and revenue.

```
from scipy.stats import pearsonr
corr, _ = pearsonr(sales_data['date_ord'], sales_data['실매출'])
print('Pearsons correlation between date and profit: %.3f' % corr)
```

```
Pearsons correlation between date and profit: -0.072
```

Apart from this, I have also calculated and visualized sales by weekday, three time periods including breakfast, lunch, and dinner, week, and month by grouping the sales data into these given datasets and visualizing these values using matplotlib.

All of the data analysis results were visualized using Tableau. I have visualized sales by different time intervals and compared the different predicted values using ARIMA, SARIMA, Linear Regression, and Random Forest.