NYIT New York Institute of Technology

Business Analytics Project

Walmart Sales Forecast - Inventory Shift Classification Model

Group 2:

Member 1

Member 2

Christopher Griffith

Member 4

BUSI 650_M01/W01

Professor: Rakesh Mittal

August 23, 2022

CONTENT

Executive summary

Based on the Kaggle Dataset for Walmart Sales between the years of 2010 and 2012, a classification model was developed to determine if a department within one of the forty-five stores within the dataset should shift their inventory to another department before a given week to increase profits. Three classification models, Decision Trees, Random Forest and Bagging, were compared based on performance metrics to determine which was suitable for this task. The features or variables used for the prediction were 1) static conditions for a store such as store, department, type and size, 2) the reporting quarter within the financial year along with the holiday calendar and 3) dynamic regional conditions such as temperature, fuel prices, unemployment rates and the consumer price index for the given week. All of these features were used to determine if a given department should be classified to downsize or remain unaffected. After reviewing the performance of each untuned and tuned classification model, it was identified that Tuned Bagging Classifier yielded the best metrics while the Tuned Decision Tree Classifier was the most consistent between the training and test datasets. Both models' performance were spectacular despite the Tuned Decision Tree lower yields.

Introduction and Description of Data

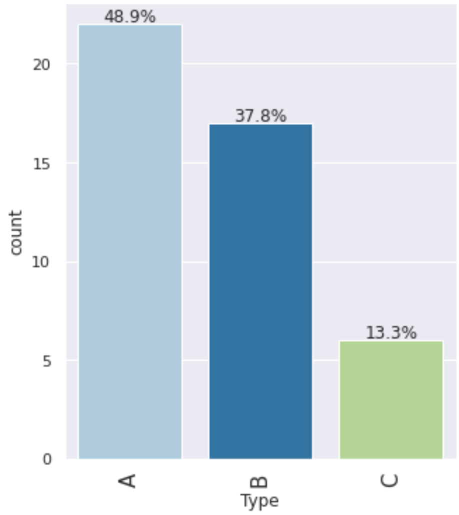
Our group will forecast Walmart's more effective ways about inventory allocation to achieve the goal of better sales. The database we chose is about Walmart's weekly sales in 45 stores from 2010 to 2012, which include the features like Temperature, CPI, Fuel Price and Unemployment Rates. Our group will increase or decrease the inventory data in a given store, and focus on the percentage shares of sales. As a result, we could predict how to shift inventory to another department in a given store to gain more profit**.**

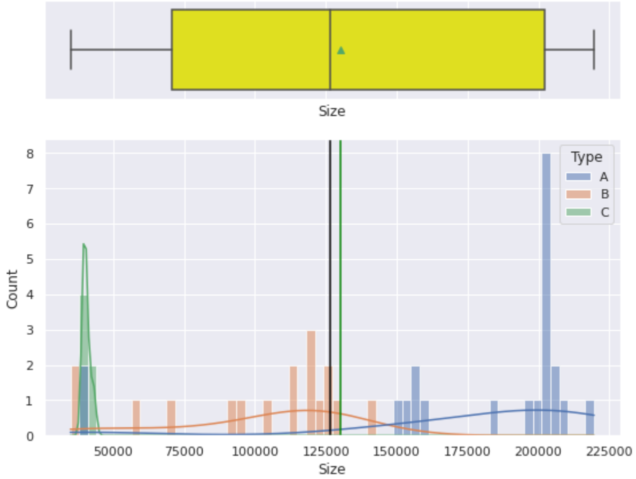
There are 45 stores in the "Store" table. The mean in this dataset is the 23 and has a size of 130,287.6 while the min store only has size of 34,875. However, the max store has the size of 219,622 which is nearly 6.3x larger than the min.

There are 421,570 records in the "Sales" table which includes items like Store, Dept, Date, Weekly_Sales and so on. From the dataset, it is obvious that the mean weekly sales are 15,981 and the min is only -4,988 while the max is 693,099. Therefore, the standard deviation of weekly sales is 22,711 which is a little high.

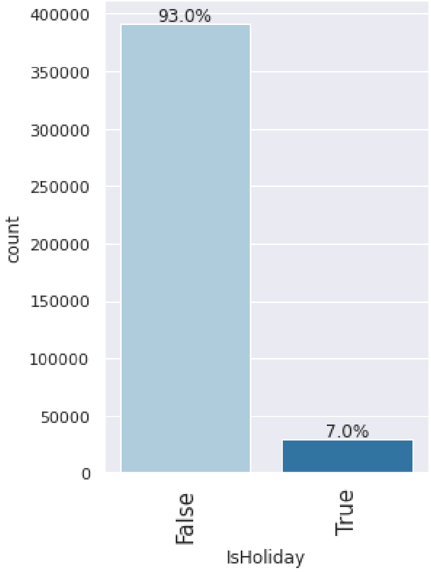
There are 8,190 records in the "Features" table which contains items like Temperature, Fuel_Price, CPI and so on. From this dataset, our group could find the relationship or impact between stores' inventory level and these features.

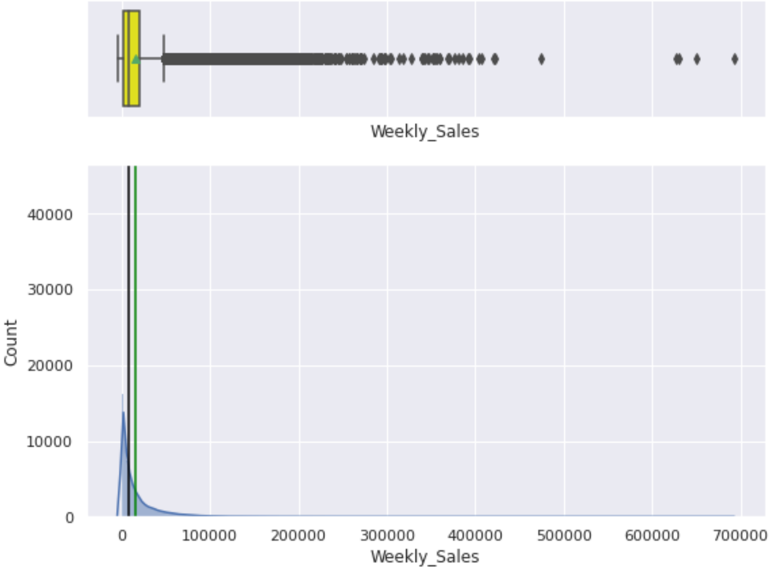
Exploratory Data Analysis





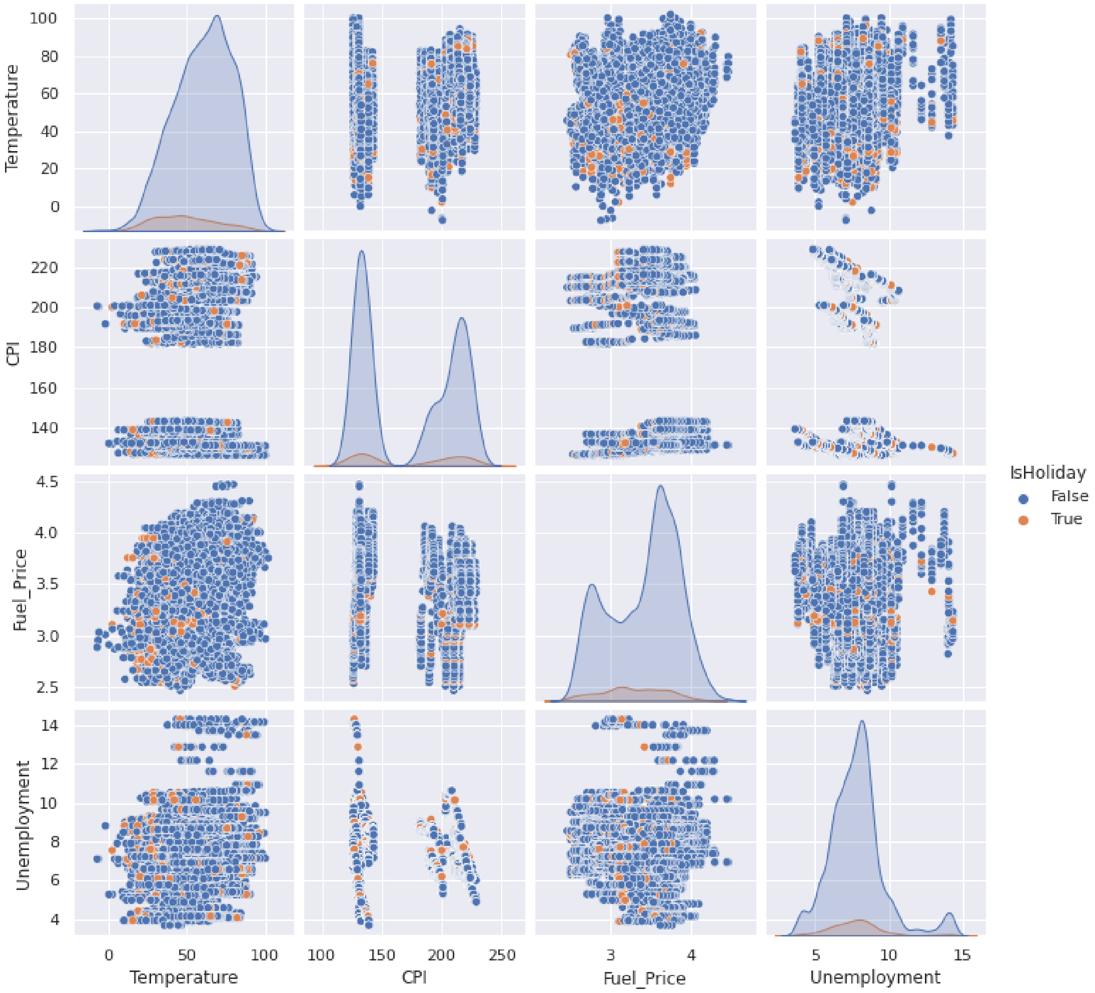
According to the size of the store, the samples (45 stores) can be divided into three types: type A, B, and C. Almost half (48.9%) of the stores belong to type A. Type A stores are larger than 150,000. The largest store size does not exceed 225,000, which is 219,622. And 37.8% of the stores belong to Type B, most of the stores in Type B are concentrated around 125,000. Type C stores account for the least, only 13.3%. This shows that among the 45 stores, the majority of stores are medium and above.





There are 421,570 records in the Sales dataset. It specifically shows the records of different departments in the 45 stores at different times. “IsHoliday” and “Weekly_Sales” are key features in this dataset. “IsHoliday” is defined as whether the week is a special holiday week, “Weekly_Sales” is the sales for the given department in the given store.

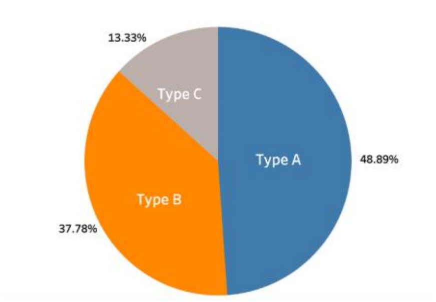
In this dataset, 93% of the weeks are not holidays, and 7% of the weeks are holidays. It can be seen from the histplot and boxplot that the average value of weekly sales is greater than the median, and the frequency distribution is skewed, that is, positively skewed.



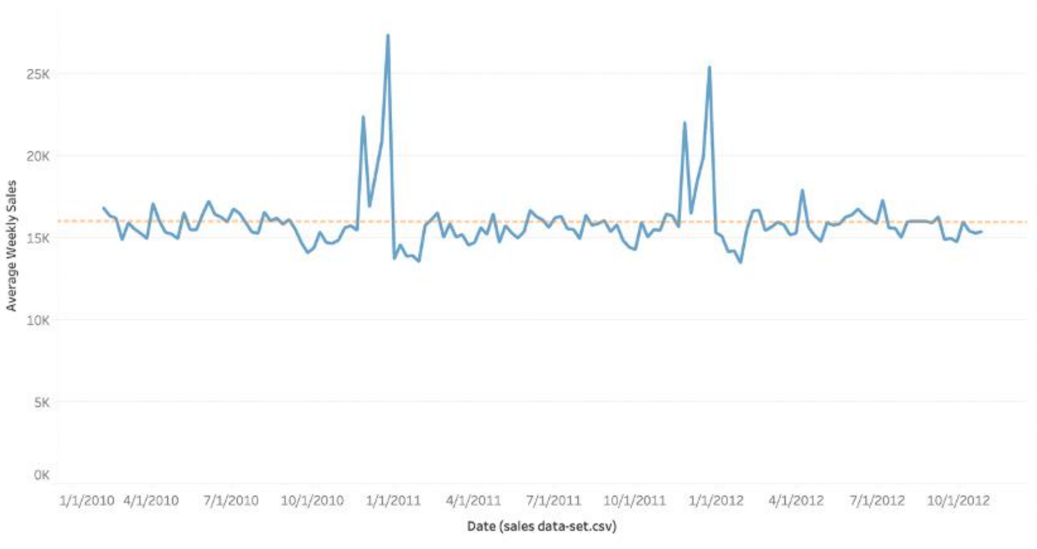
The pair plot (from the Feature dataset) of Temperature seems to present a normal distribution. Other features are the presentation of bimodal. In the exploratory analysis of the Feature dataset, we use holidays as the classification criteria. But holidays are much less than non-holidays, so the distribution of the two colors in the counts of temperature, CPI, Fuel_Price, and Unemployment shows significant differences, which is a reasonable result.

Initial Business Insight

Before processing the three datasets, there is a limit to the information we can analyze. Since we need to extract and integrate for the data that can be studied. However, from a preliminary understanding, we can draw general conclusions about the types and characteristics of the data.



In the "Store" dataset, we have a sense of the size of the 45 stores sampled by Walmart, of which almost half are Type A stores, i.e., stores over 150k in size. This means that type A stores are more popular than B and C types, and as a matter of common sense, larger stores with higher customer carrying capacity and a wider variety of products are likely to outperform smaller stores in terms of sales. We can further verify this with the data from the "Sales" dataset in the subsequent analysis if necessary. The "Sales" dataset contains the crucial variable "sales", which is one of the important objects of our study. What can be analyzed is that, for a given time frame, there are lows in January in 2011 and 2012, and peaks and sub-peaks in 2010 and 2011 between the end of November and the end of December. During the period from February to October each year, weekly sales remain almost constant at around 15k.

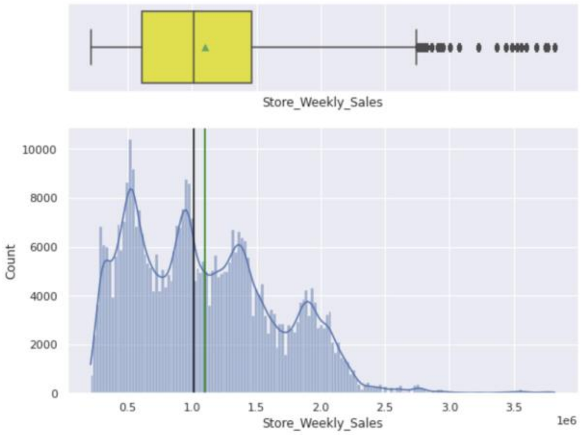


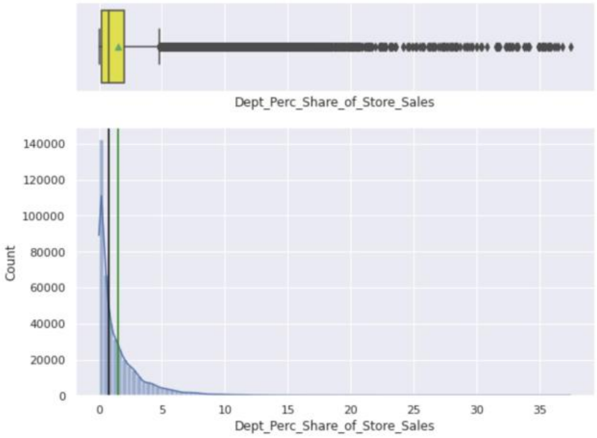
The "Feature" dataset has data on variables such as temperature, fuel prices, consumer price index, and unemployment, and whether they are likely to have an impact on sales per store is unknown to us in our initial exploratory analysis. We need to integrate the data set for analysis. Also based on EDA and Correlation studies, columns with weak relationships to the target columns can be removed.

Development of the Merged Dataset

Since we have three separate datasets, it would be useful to have all of the variables describing the sales for Walmart within one dataframe. More insights about how the stores and departments reported their sales could be extracted if the datasets were merged into one. Using the primary keys in each table, we were able to merge the datasets while preserving the features unique to each table. The merge dataset of the three tables now contains the size of the store, the type (A, B or C) of the store and the regional conditions along with the sales figures reported for each record. Now we are able to provide more insights on how the conditions such as Temperature, CPI, Fuel Price, and Unemployment Rates relate to the stores and departments.

Total Store Sales, Department Contribution and the Downsize Class





We were able to determine how much sales a store collectively earned compared to the percentage of contribution of each department from each store. More than half of the sales records reported sales less than the mean $1.1 million. Based on this distribution, more than half of the sales reported by the departments contributed less than 1% of total store sales to their home store for a given week. This revelation provided a means to create a classification model aimed at identifying which departments could potentially contribute less than half of the reporting departments. Using a 0.75% threshold, the class downsize was imposed on all of the records indicating if a department contributed less than .75% to the total sales of a store for a given week. This class will be predicted using a classification model without any sales metrics from the stores or departments. The models will rely solely on the 1) static conditions for a store such as store, department, type and size, 2) the reporting quarter within the financial year along with the holiday calendar and 3) the dynamic regional conditions such as temperature, fuel prices, unemployment rates and the consumer price index for the given week.

Research Method

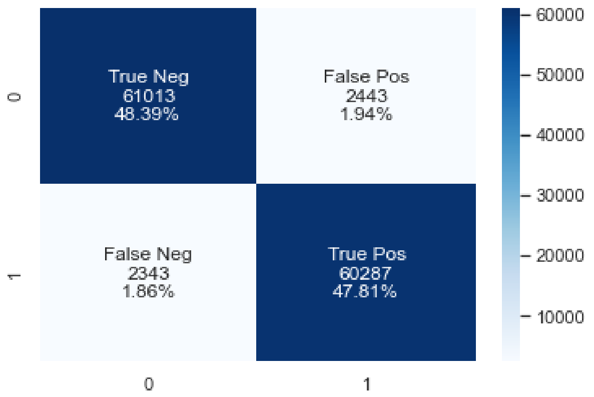
To achieve the goal of better sales by inventory allocation, our group will use three research methods: decision tree classification, random forest classification and bagging classification.

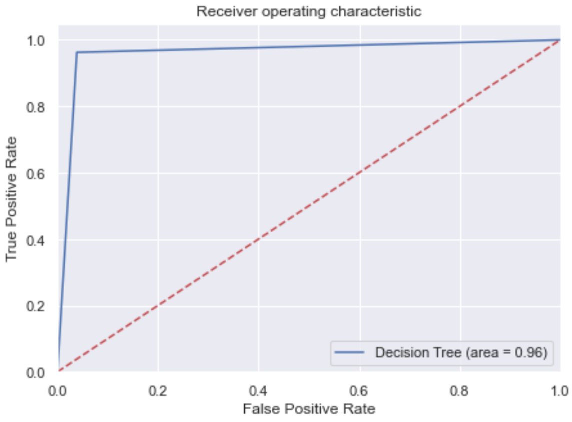
Decision tree classification is a graphical representation of all the possible solutions to a decision that is based on a certain condition. It is a popular and effective supervised learning technique for classification problems that works well with both categorical and continuous variables.

Random forest is a supervised machine learning algorithm that is used widely in classification and regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

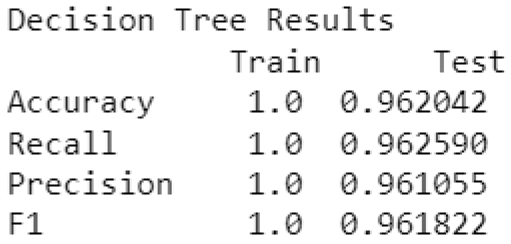
Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions.

Classification Model and Results



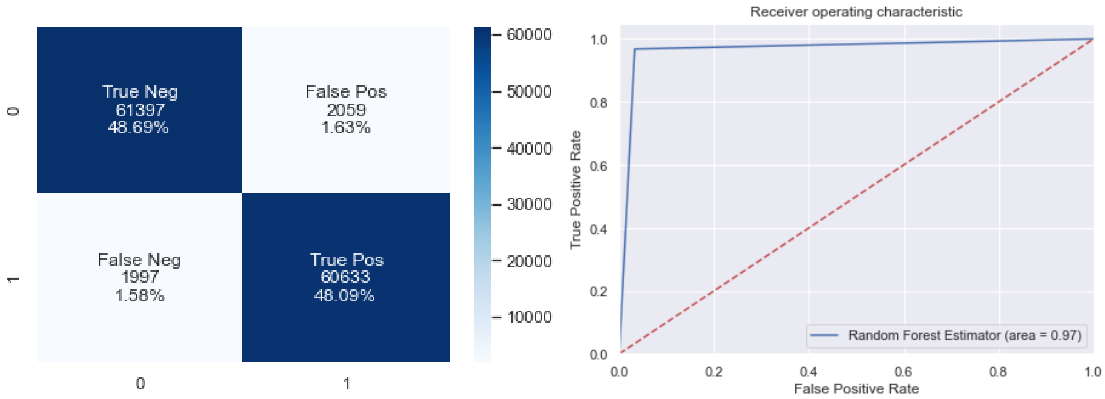


In our first classification model, we started off with the Decision Tree classifier. Our prime goal was to maximize the F1 score because we wanted to make an optimal decision from an inventory perspective. We wanted to make such a decision to transfer the inventory in such a given time before a week to maximize our profits. In our untuned decision tree, the training data overfits the model; however, our test data results performed well.

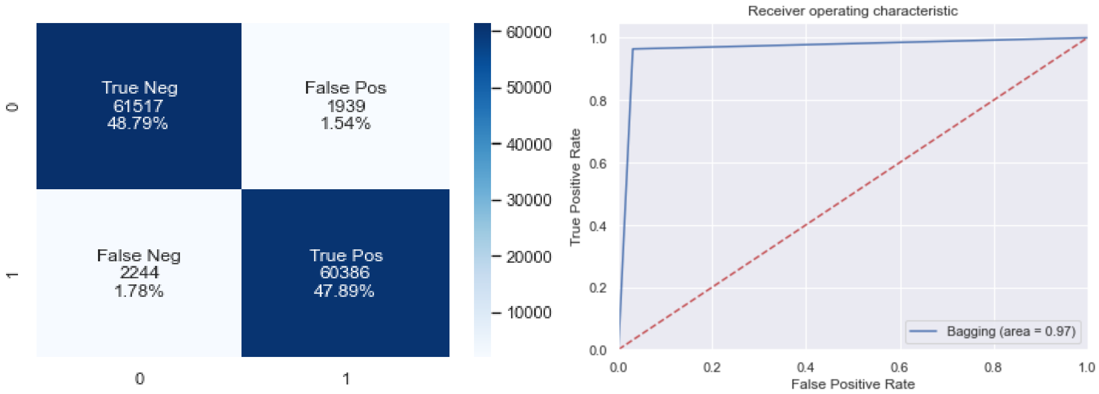


```
Decision Tree Results
              Train      Test
Accuracy        1.0  0.962042
Recall          1.0  0.962590
Precision       1.0  0.961055
F1              1.0  0.961822
```

As we can see in the result, our train data overfits as it has 1.0 in all four metrics and the test data performed well. Size, Type B (Stores), Temperature, CPI and Fuel Price were the top 5 most important features in our Decision Tree.



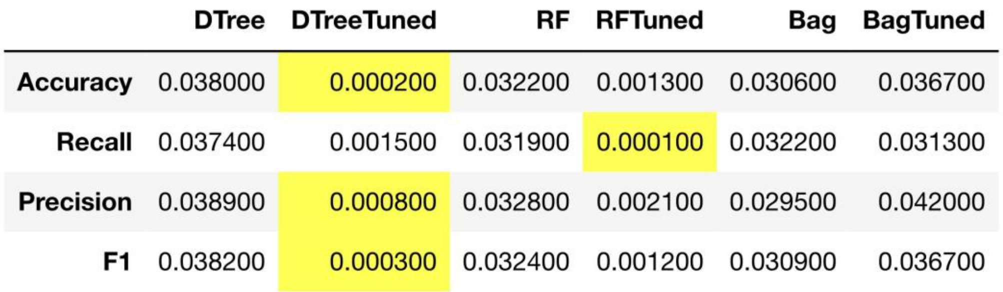
After the decision tree classifier, we needed a more focused model which is the Random forest classifier. This model utilizes many decision trees of different samples and takes the majority vote for classification. After classification, the results were quite similar as we had in the Decision tree. The train data overfits and test data performed well for the Random Forest classifier. The feature importance of the Random forest classifier was Size, CPI, Temperature, Fuel Price and Unemployment Rates. These were the five important features which can help make the decision on whether or not we should shift our inventory to the respective department and how much profit we can make in a period.



The bagging classifier can help us classify our data that fits the best and can run the test on each random variable of the all original department data set and then aggregate their individual predictions. This classifier helped us to form a prediction as it has the better performance compared to the other two. We still have a sign of overfitting and yet again our test data performed well. The bagging classifier can predict the departments which need to downsize their inventory; however, it can misclassify the departments where the inventory allocation should transfer. At Least we can satisfy our focused goal; however, these results can lead to disrupt the supply and respective department sales.

The tuned models, as we all discussed in class, can get better results and reliability, where in our case we had good results in our last bagging classifier where our F1 score was utmost we had a slight chance of misclassification in prediction or results. The tuned decision tree and random forest models initially lost performance but at the same time gained reliability and in the results the tuned decision tree and random forest classifiers did not overfit as compared to their untuned models. After tuning, it appears that although the Tuned Bagging Classifier had the best overall performance, the Tuned Decision Tree was the most consistent between the Training and Test data. Despite the lower performance, the Tuned Decision Tree would produce results consistent with the training data, our F1 score was the prime metric since we would like to minimize both false positives and false negatives. This translates to minimizing the misclassification of departments that should NOT downsize and minimizing the misclassification of departments who should downsize.

Conclusion



	DTree	DTreeTuned	RF	RFTuned	Bag	BagTuned
Accuracy	0.038000	0.000200	0.032200	0.001300	0.030600	0.036700
Recall	0.037400	0.001500	0.031900	0.000100	0.032200	0.031300
Precision	0.038900	0.000800	0.032800	0.002100	0.029500	0.042000
F1	0.038200	0.000300	0.032400	0.001200	0.030900	0.036700

The best performing model, the Tuned Bagging Classifier, was impressive despite the slight overfitting from the training data. However, we believe the safest and most reliable was the Tuned Decision Tree due to how close the metrics were between the training and test datasets. In the delta performance metrics chart above, the tuned decision tree has the smallest change between the training and test data. Based on the results, it appears we are able to predict with great accuracy, precision, recall and f1 score, if a department within a Walmart store should shift its inventory to increase the sales or revenue of a given store. This predictive model could potentially save Walmart millions of dollars in unsold inventory while increasing sales overall for its stores.

References

Walmart Sales Forecast. (n.d.). www.kaggle.com. Retrieved August 23, 2022, from https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast?select=features.csv

Retail Data Analytics. (n.d.). Www.kaggle.com. Retrieved August 23, 2022, from https://www.kaggle.com/datasets/manjeetsingh/retaildataset?resource=download&select=sales+data-set.csv