Group 8

Michel Leonard

STAT GR5293: Topics in Modern Statistics, Applied Machine Learning in Financial Modeling

17 May 2019

**Predicting Directionality Of Movement Of Automotive Index Using Historical Data And Financials**

**ABSTRACT**

The automobile industry exhibits seasonal patterns, and in this paper we demonstrate the value add of prediction performance improvements in terms of accuracy over using conventional linear models such as Logistic Regression by Machine Learning techniques such as Random Forests, Support Vector Machines, and Gradient Boosted Trees. We observe that the Machine Learning techniques significantly outperform Linear Models in similar test conditions.

# **INTRODUCTION**

# BUSINESS UNDERSTANDING

Cyclical patterns in the Automotive Industry are a well-studied topic, and of great interest to traders who need signals on the market to identify trends. In this paper, we seek to leverage information from cyclical patterns using Machine Learning Techniques to predict the direction of movement of the Seasonally Adjusted Automotive Registrations Total over Time (SAARTOT). Prediction of movement of SAARTOT index is a useful indicator for identifying trend reversal patterns, and can thus be used as a feature for regime switching models.

The automotive industry has been historically characterized by long product development cycles, typically varying from one year to 5 years in advance. There are three major factors influencing the growth of Automotive Industry: Innovation, Marketing, and Macro factors. For innovation, most of the traditional auto firms are seeking new ways to use clean energy. However, as there are far many factors to consider in the innovation cycle, speed and quality, innovation itself is difficult to interpret in a Quantitative setting. For marketing, the competition among firms is strong while the extent of product differentiation is comparatively small. Hence, we believe that the marketing effect might be small. Also, as marketing typically has more of a short term impact than a long term impact, we focus on Macro factors and cyclical trends for predicting the possibility of growth of Automotive Industry.

In this paper, we compare the predictive power of Machine Learning Techniques such as Support Vector Machines, Decision Trees, Gradient Boosted Trees etc. versus a Linear Model (Logistic Regression) based on in sample/out sample performance using metrics such Area Under Curve Receiver Operating Characteristics (AUC ROC), Average Precision and Accuracy. We observe that Machine Learning Techniques substantially outperform a Linear Model in terms of accuracy and AUC ROC.

We source our Data from Bloomberg in a month over month format, and use the Seasonally Adjusted Registrations Total over Time as the target for our predictive models. We obtain data for indicators such as Gross Domestic Product, Consumer Price Index, Producer Price Index, Purchasing Managers Index, M1, M2, National Gasoline Price etc. in addition to the sales data for 20 top brands in the automotive market in the United States which were used as covariates in our models. We establish the presence of cyclical patterns in the data visually as described in figure 6.

We model the problem as a Classification task, where the objective is to predict if the Automotive Industry is going up or down in the next time frame. We use Machine learning techniques because of the empirically validated performance improvements we observed over Linear Models.

We perform feature selection manually, as well as using algorithms such as Forward Selection and Backward Elimination to identify important features, in addition to univariate and bivariate analysis. We also discuss the problem of outlier detection, and outlier removal in the Data Understanding section. We discuss in detail about the need for using Forward-Backward value filling algorithms for handling missing data, and the value adds of using dimensionality reduction techniques such as Principal Component Analysis etc.

Finally, we discuss our implementation in detail, and analyze the results to show empirically that Machine Learning techniques substantially outperform linear models (Logistic Regression) in terms of out sample predictive performance using metrics such as Accuracy, Precision, F1 score and AUC ROC.

# **DATA UNDERSTANDING**

***Data Description***

Our main target variable is SAARTOT index (Seasonally Adjusted Automotive Registrations over Time), which captures the information about automotive sales patterns in the United States of America. SAARTOT index essentially provides an adjusted value for predicted annual automotive registration sales using historical averages for Monthly registrations data. We quantize the SAARTOT index using lagged data to include information on whether the index is going up or down compared to the last data point. We quantize the response by the following transformation

The covariates for the models are sampled from monthly sales data for top 20 manufacturers in the United States of America, and also contain other economic factors such as GDP and etc. The full list of features used in our models is described in Table 1.

**TABLE 1. Feature Description**

|  |  |  |
| --- | --- | --- |
| Variable number | Variable name | Description |
| 1-20 | Sales data | Top 20 auto brands in the US, including Audi, BMW, Mini, Mercedes Benz, Mitsubishi Fuso, FCA, Honda Group, Hyundai, Isuzu, Jaguar, Land Rover, Kia, Mazda, Mitsubishi, Nissan, Porsche, Subaru, Tesla, Toyota Group, Volkswagen, Volvo |
| 21 | Real GDP | An inflation-adjusted measure that reflects the value of all goods and services produced by the US |
| 22 | US auto production | Domestic auto production (defined as all auto-assembled in the US) |
| 23 | US auto consumer SA | Fed Autos & Trucks consumer SA |
| 24 | US Auto parts | Fed auto parts & allied goods SA |
| 25 | National average gasoline price | Mid-price of national average gasoline prices (regular unleaded) |
| 26 | PI US URBAN Consumers MoM SA | Price index for urban consumers SA in the US |
| 27 | PPI US finished goods SA | The average change over time in the selling prices received by domestic producers for their output |

Our data was sourced from Bloomberg, widely considered as one of the cleanest and most reliable vendor for financial data. The timeline for the data is between January 2000 and December 2019. We sample the data on a monthly basis and thus obtain 240 data points for our analysis. The response variable is binary, and the independent features are continuous numerically valued variables.

In our experiments described in detail in the experiments and model selection section, we observe that several indicators such as M1 and M2 index have no tangible impact on model performance, and hence we discard them in our final model. Table 2 describes the features which were used in the experimental phase, but were discarded in the final model.

**TABLE 2. Dropped features**

|  |  |  |
| --- | --- | --- |
| Variable number | Variable name | Description |
| 1 | PMI ISM manufacturing SA | The measure of the prevailing direction of economic trends in manufacturing, based on a monthly survey of supply chain managers across 19 industries, covering both upstream and downstream activity |
| 2 | M1 | Money supply that includes physical currency and coin, demand deposits, travelers’ checks, other checkable deposits, and negotiable order of withdrawal (NOW) accounts. |
| 3 | M2 | A calculation of the money supply that includes all elements of M1 as well as savings deposits, money market securities, mutual funds, and other time deposits. |

***Data Understanding***

The selected top 20 brands capture roughly 93% of total sales in the US. Hence we assume that most of the information about automotive sales is captured by the top brands. The automotive sales data contains data about light vehicles, cars and light trucks, and combined them to obtain the total automotive sales for each manufacturer. This was done to make the data more consistent with the SAARTOT index, and also to reduce the dimensionality of the data. The aggregated sales data might have significant value add due to the presence of cyclical patterns.

GDP captures the total spending potential of the economy, and hence a useful indicator to test the health of the economy. The higher the spending potential, the better equipped the target consumer is to purchase a non-essential depreciating commodity like an automobile. Hence, there is a direct relationship between GDP and Automobile sales.

Assembly of auto parts plays an important role in the car production process. Domestic as well as International brands regularly produce most of their car parts domestically, therefore we expect that the availability of car parts and their materials would have an impact on the prices of the cars and hence overall to the total sales.

Gasoline is a recurring expense on the automobiles, and consumers are very price sensitive to recurring expenses as they add up significantly over time. Hence, we establish a negative correlation between gasoline prices and automobile sales.

PPI refers to the money that domestic producer received for their output. The amount of money could affect the revenue of domestic producers and hence affect their operating level. When the revenue gets lower, auto firms may change their plans for the production level. As for PMI, the auto industry is one of the 18 industries included. Consequently, we believe that the economic trend in the manufacturing industries could have an effect on the auto industry.

M1 and M2 describe the money supply in the US, and it may affect the investment, supply, and demand of the auto industry. If the money supply is abundant in the auto industry, this may stimulate the growth of the whole industry and encourage consumer confidence in purchasing cars.

In our final model, we observe that most of the variables we included are significant, while M1, M2 and PMI are not significant. This could be explained as below:

M1 and M2 refer to the money supply in the US. We consider they are affecting the investment in the auto industry. However, the stimulation of M1 and M2 could be slow based on three factors. First, the auto industry is a relatively mature industry, indicating a rather stable demand and supply. Hence the additional investment in the auto industry might have little effect in the development of the whole industry. Second, production machines usually have a relatively long service life. Although there is additional investment, auto firms usually will not phase out old machines when they can still work in the factory. Hence, the renew of production machines can usually be slow and hence the production level could hardly improve within a short time. Finally, auto consumers are usually taking loans when purchasing a new car. Larger money supply in the investment of the auto industry will not directly lead to an increase in the available money that consumers have. Hence, they might have little influence on the sales of the auto industry. Finally, we see in the later sections that there are many missing values in the data for this feature, and hence maybe the model is unable to fully leverage all the information captured by this index.

As for PMI, there are in total 18 industries included in this index, and the auto industry is just one of them, taking up less than 20 percent of the total weight. Hence, we believe that this might be the reason why they are contributing relatively little to the auto industry.

***Sampling Method***

The data with which we are dealing has an implicit time series structure to it, hence a fixed sampling method is the most appropriate approach to generate the in/out sample. We obtain 240 observations for our data, and we split it into the first 190 for the insample (training dataset) and the latter 50 for the outsample (test dataset). This is done to ensure that there is no data leakage from future, by ensuring that the model learns to generalize well. We measure the performance of the classification models we build on the basis of the outsample performance.

***Data Manipulation***

The SAARTOT index obtained from Bloomberg are numerical and on a monthly basis. However, in this project, we are trying to predict the direction of movement of this index. Therefore, we first find the monthly return of SAARTOT index and find the sign of the monthly return

Also, as we are trying to forecast the movement of the index, we shift the response variable toward the previous timestamp by 1 unit. In this case, we are using the factors at time to forecast the movement of SAARTOT index at time.

Bloomberg is known to contain some of the most clean and reliable financial data, some of the features that we obtained had a large number of missing values. Removing missing values on the basis of Complete Cases would result in very inferior models as Machine Learning techniques typically require as much clean data as possible. Complete case can be defined as the case where all the values for each feature is available for a data point, while available case analysis works on the basis of univariate information about each feature. Dropping data points on the basis of Complete case analysis leads to a drastic decrease in the number of samples for fitting and testing data, and we found that it does not yield the most optimal results.

Filling in with univariate information simply would lead to leakage of data from future, because of the time series aspect of data. The best approach would be fill in missing values using the values from the last available datapoint for the feature. If the last known value is unknown, then we can interpolate the future value to the missing value to make it consistent. This algorithm is known as Forward-Backward filling algorithm.

The forward backward filling algorithm first fills the missing values using information from the previous state unless the information is not present. For the entries for which the imputation for missing value couldn’t be determined, the Forward Backward algorithm then fills the missing value with the next knowb value in the future. Filling the missing data in a univariate fashion allows us to essentially do an available case analysis on the data, and hence we are able to fit our models on a larger dataset. We further perform outlier detection and removal using univariate tests, and also remove outliers on the basis of visual inspection over variable relationships.

Sales from top 20 manufacturers has significant collinearity due to the nature of the automotive industry, hence we try to perform Principal Component Analysis on our data to reduce dimensionality and remove collinearity. PCA essentially rotates the data in the spatial space in such a way that features are orthogonal and hence have no strong correlation between them. PCA can be easily imagined as a rotation which tries to create a multivariate gaussian with no covariance between the dimensions. The dimensions with the highest variance are called the principal components, and higher the variance in each dimension the more the information that dimension or principal component captures. Selecting the dimensions with the maximum variance allows for capturing the maximum number of variance in the data, and thus we can discard the dimensions with lower dimensions thus performing dimensionality reduction. This method causes the data to lose interpretability as the new features are often a combination of the raw features, and by performing dimensionality reduction we are often unable to recover the exact relation of data that was discarded in terms of loss. We also observe no significant value add in terms of performance improvements in Machine Learning techniques hence we chose to not include it in our final model.

We perform feature selection manually, and remove all features that have no significant value add as described in Table 2. We perform feature selection to reduce training/inference time, avoid curse of dimensionality and to enhance generalizability of our models by constraining the ability to overfit on noise. Feature selection has an added benefit of making the model more interpretable by simplifying the features.

# **MODEL UNDERSTANDING**

***Model Goal***

The objective of the model is to predict whether Automotive Registrations for the next month are going to go up or down. We model it as a Classification problem with a Time Series aspect to it.

***Model Selection***

We model our problem as a classification problem, and hence we choose Logistic Regression as our baseline linear model for performance comparisons with Machine Learning techniques such as Random Forests, Gradient Boosted Trees and Support Vector Machines.

Logistic Regression uses a Logit link function to map responses from a continuous real values scale to a unit scale between 0 and 1. The logit function can be interpreted as a probability distribution function, and hence we can infer the probability of each class on the basis of closeness to either zero or one. The continuous real valued responses are generated using Linear Regression, and we perform transformation using logit link function to obtain the class confidence probability for each class. Each class here represents if the automotive index is going up or down in the next time stamp.

We also perform experiments using Machine Learning techniques such as Support Vector Machines, Random Forest Classifiers and Gradient Boosted Trees.

Support Vector Machines perform classification by projecting the data into a n-dimensional hermitian space, in which n is defined by the Kernel function. A kernel function maps the data from the original representation to the higher dimensional representation for classification. Further, a Support Vector Machine learns to classify points into classes on the basis of Hinge Loss function which tries to maximize the distance between the support vectors of the classes. A support vector is essentially a data point mapped in the higher order space which is situated at the periphery of the class representation. Projecting and classifying data in a higher dimension space has a high computational overhead, and it is circumvented by using a Kernel Trick which allows for only computing distance between two data points in the higher order space and not the exact location itself. We use the kernel trick in Support Vector machines by using a Radial Bias Kernel function which uses Gaussian distribution to map the data into a infinite dimensional space, which might have been computationally infeasible without the kernel trick.

Furthermore, Logistic Regression implicitly provides class confidence probability values but Support Vector machines generally provide hard class labels. The SVM’s can be made to predict probability values using distance measures from the decision hyperplane constructed by the classifier using hinge loss.

We also perform ensembling over decision tree classifiers using Random forests and Gradient boosted trees. Decision trees essentially perform a if-else conditional statement on the basis of values in each feature. The decision to set the boundary for if-else is determined by loss functions such as Gini, Entropy etc. The loss functions in decision trees reduce randomness in the response labels at each step, in the end stopping when there is only one class predicted by a condition. Decision trees can have infinite number of conditional statements if left unconstrained, and this might lead to overfitting the data and the noise in the data. We can circumvent this issue by constraining the number of conditional statements, and also by constraining the data on which the models are trained.

Ensembling methods work like a democracy in which each classifier has a vote, and the class with the majority vote is predicted as the class of the response variable. Ensembling essentially tries to balance diversity and accuracy in its models, and the diversity can be enforced by constraining the model parameters or the data sampling procedure.

Random Forest method builds decision trees on the basis of randomly sampled subset of the features and data points, and then averages the predictions to generate the class of the response variable. Gradient boosting builds trees greedily by optimizing trees to predict better on misclassified samples from the previous decision trees. Gradient Boosted trees are known to fit the data better than Random Forests, but are also prone to overfitting on random noise. We observe in the Experiment and Model selection section that Random Forest Classifier and Gradient Boosted Tree Classifier have the best in sample/out sample performance when comparing different classification models.

Random forest classifiers are essentially an extension of decision trees. It constructs multiple decision trees using a random subset of data and a random subset of variables and merges them together to get a more accurate and stable prediction with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

***Feature Relationship***

After exploring the data, we notice that there is a correlation between automotive sales for the current & previous months and the future direction of movement of Automotive sales. Besides, indicators such as GDP, Gasoline Prices have a strong correlation with the target variable.

First, we try to understand the correlation between automotive sales for the current and previous months and the future direction of movement of automotive sales. Usually, people’s passion of purchasing a car is affected by the economy and economy is rather steady in a short time. Thus, based on common sense, automotive sales should not fluctuate in a big range.

Also, indicators such as GDP, Gasoline Prices have a strong correlation with the target variable. As mentioned before, automotive sales are relatively steady in a short time and their movement is determined by economy, including GDP, CPI, money supply and so on. If the economy is great, people receive higher disposable income and have extra money to buy cars. Besides, automotive sales are influenced by the price of cars substitute. If gasoline prices go up, the cost of owning a car rises, less people tend to buy a car, and sales decreased.

Comparing the influence from the correlations mentioned above, the biggest impact is coming from the sales data for the previous timestep, but we see reasonable improvement using other indicators such as GDP, us auto production, auto consumer SA, auto parts, national average gasoline price.

# **Descriptive Statistics**

In this section, we perform visualizations and infer relations between the different features in the data.

***Univariate Analysis***

Using Univariate analysis, we observe the presence of many missing values in the data. We plot the distributions of different combinations of features in Figure 1 to see if the data was missing completely at random, or was there a trend in the missing values on the basis of auxiliary information. We observe that the data was missing completely at random, and due to the Time series aspect of our data we chose to use Forward Backward filling algorithm to fill the missing values.

Further, we also observe that the number of outliers are rather limited, and we remove them using univariate tests such as removing on the basis of standard deviations from normal.

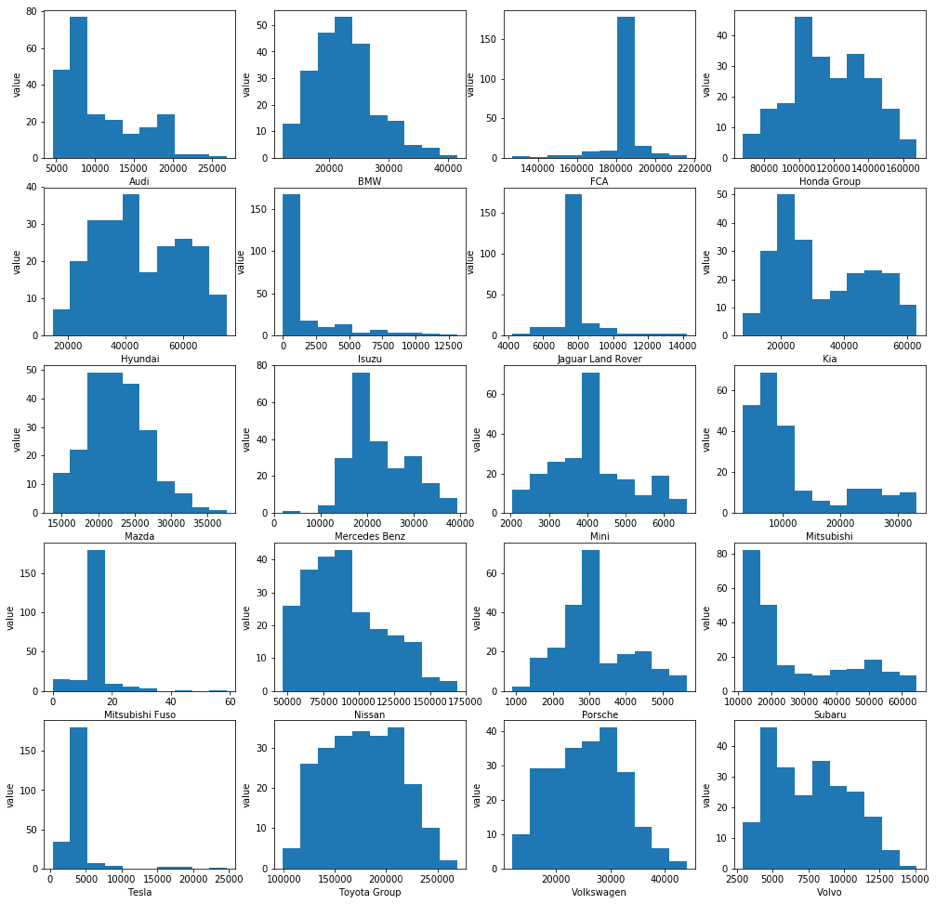


FIGURE 1. Univariate Analysis

We also observe that certain features such as M1 and M2 index have a higher rate of missing values (greater than 50% of all data points) than other features. This might be the reason for poor value add of M1 and M2 index to our prediction models. Due to the lack of value add, and the role of increasing model complexity, we chose to remove such features with high rate of missing values in our final model.

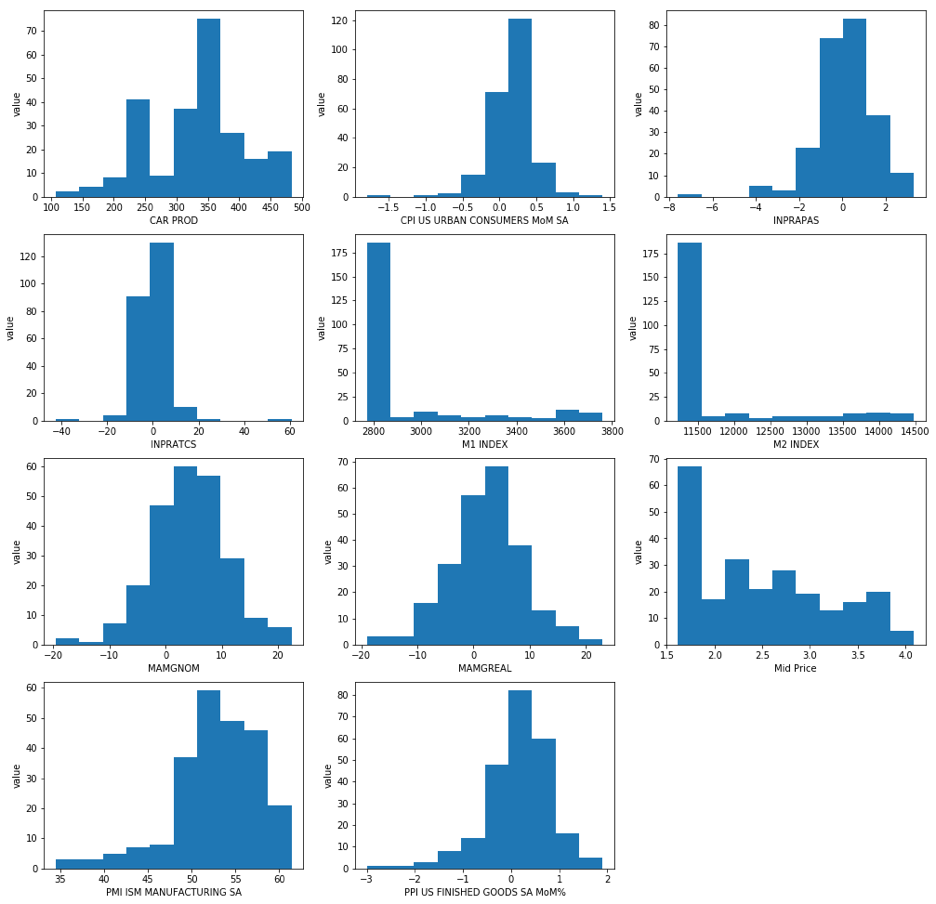


FIGURE 2. Feature Distribution after Removing Features with a large number of missing values

To get a better idea about the features, we plot the features using boxplots in figure 3. Box Plots can help identify the outliers in features. The circles in each plot refers to an outlier in the data according to univariate tests on the basis of interquartile ranges. An interquartile range is defined as the range of values in a feature spanned by the 75th percentile of value and 25th percentile of values. We claim that a value is an outlier if it exceeds 1.5 times the Interquartile Range from 75th or 25th percentile whichever is closer. We also observe that removing outliers doesn't have a significant impact on our model performance, and we infer that one plausible reason might be because some of the values aren’t inherently outliers and are rather natural.

Further, we can also infer the presence of a large number of missing values in the data on the basis of absence of the box in the boxplots. If there is no box in the box plot, we can infer that the particular feature has a large number of missing values.

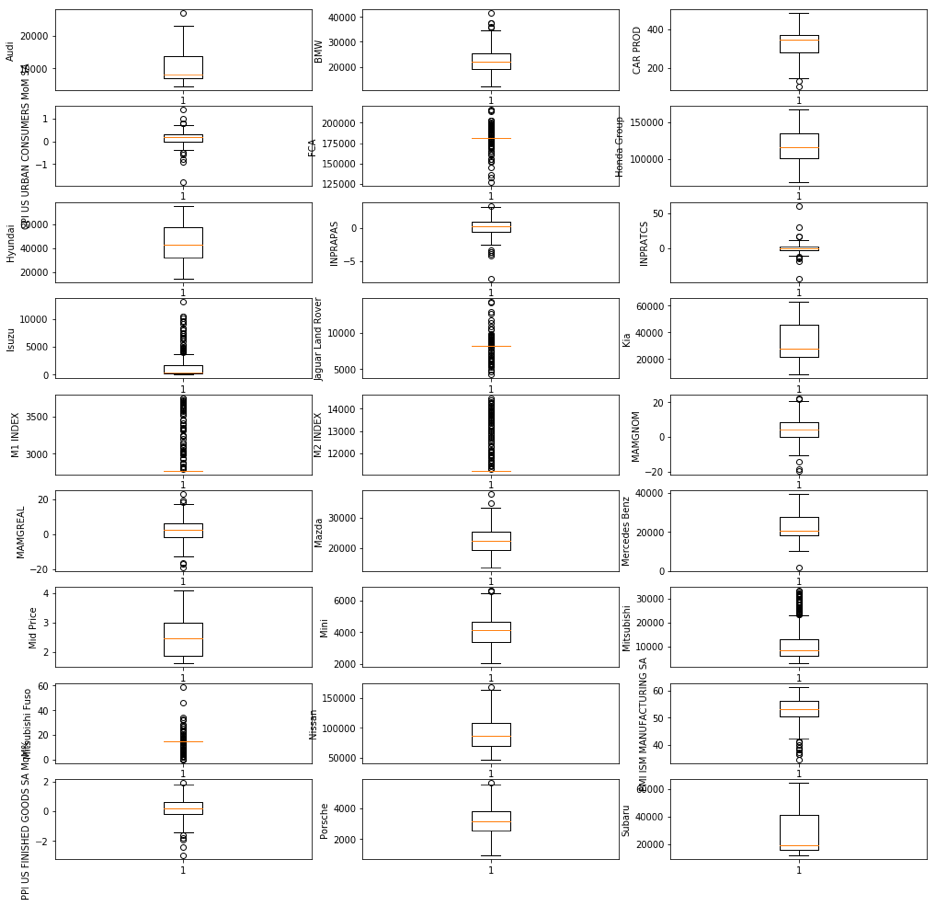


FIGURE 3. Boxplot of Features

***Bivariate Analysis***

We plot scatter plots to visually identify, establish and infer relationships between the target and the covariate variables. The response variable is binary, and all of the values are situated spatially at the ends of the graphs.

In our analysis, we observe that most features have no significant direct relationship with the response variable, in effect meaning that a change in the feature cannot be directly attributed to a change in the response variable in most cases.

In our best performing models, we observe that features and have significant impact on our model performance, and we can see relatively direct relationship betweenand the target response in Figure 4. Through our bivariate analysis, we couldn't establish a direct visual relationship betweenand the target variable.

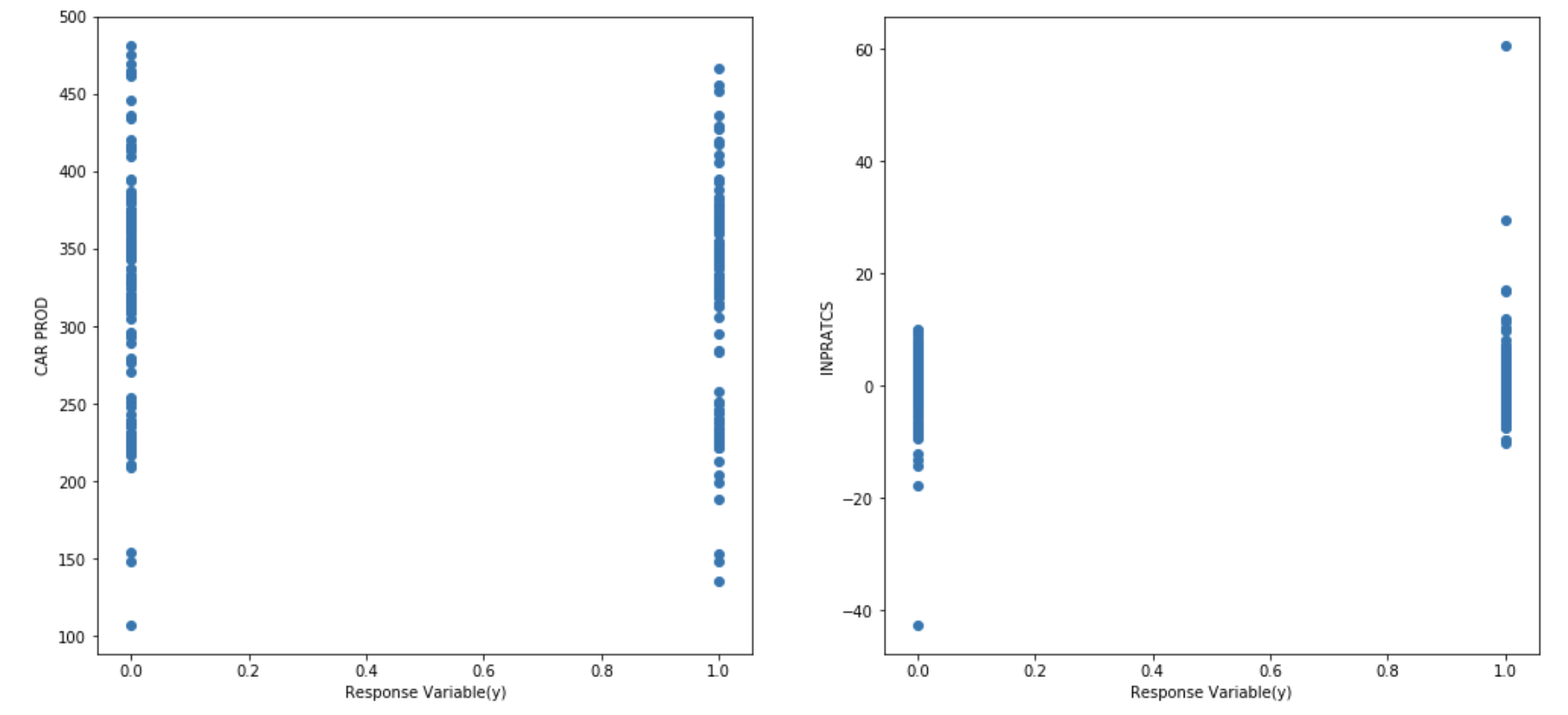


FIGURE 4. Relationship between, and the Target Response

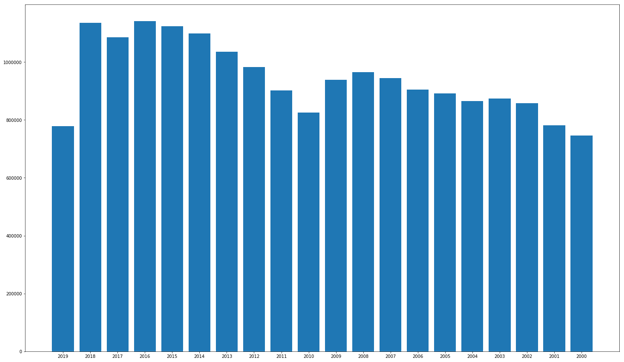


FIGURE 5. Cumulative Sales Data for the top 20 Manufacturers Per Year

We also analyzed the overall trend of sales data over all the years. We observed that for some manufactures, auto sales increased, and for some, it decreased, an insight indicating that the movement cannot be derived directly. We also looked at the overall auto sales across the years. We found out the it is gradually increasing in most of the years from which we can derived that movement is more likely to be positive in upcoming years than the previous years. However, as movement is affected by the contribution of the many factors, not just total sales data, so we expect model to learn more complicated pattern.

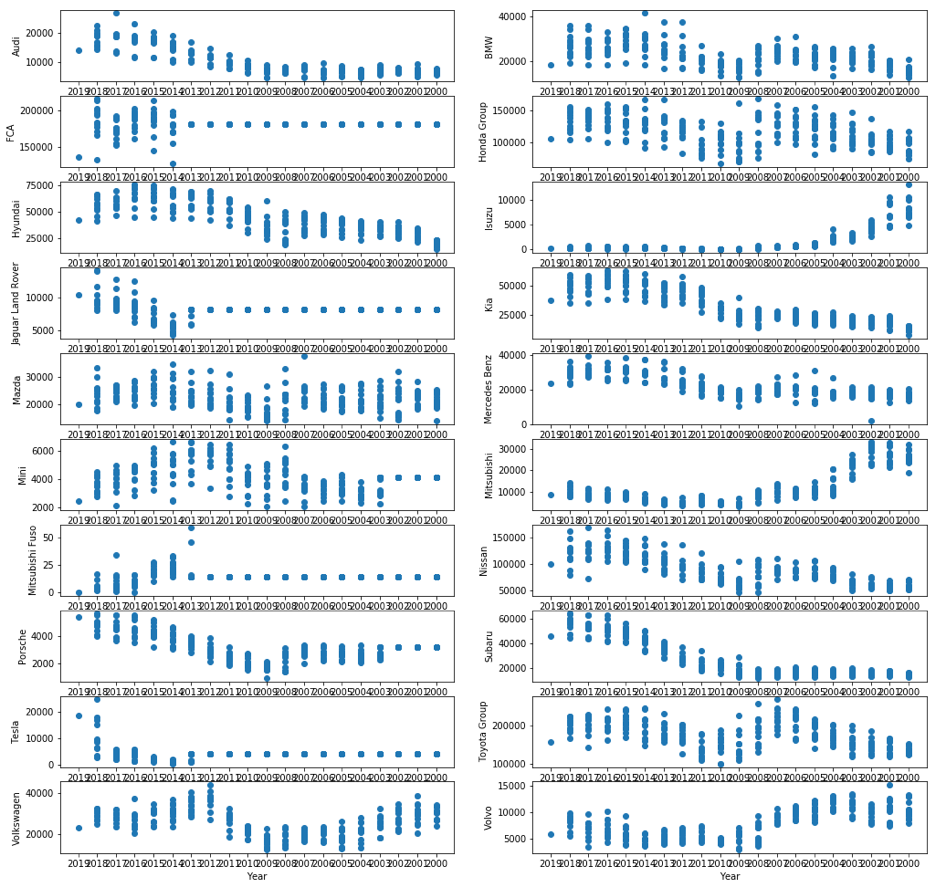


FIGURE 6. Analysis of Overall Trend of Sales Data over all the Years

# **PERFORMANCE METRICS**

We model our problem as a classification task, and we observe that the two classes defined in our problem statement are balanced in terms of support (number of data points with the target labels as the class) and hence metrics such as Accuracy, Area Under Curve Receiver Operating Characteristics are ideal for measuring the predictive performance of our model.

Accuracy measures the number of correctly classified samples made by the predictor over the total number of sample data points.. Accuracy is used to check the overall effectiveness of performance of the model over classification, but it is effective only when the classes are balanced.

Sometimes, we prefer the performance of one class over other, such as we might be more interested if the automotive index will go up than if it would go down. For this, class specific metrics such as Precision, Recall, Sensitivity, Specificity, F1 score etc. might be more appropriate.

Furthermore, typically Accuracy and Class specific metrics are used on probabilistic values which are binarized with a threshold, essentially values over a particular threshold are termed as belonging to class 1 and class 0 otherwise. We might be interested in how the classifier behaves for different probability threshold values, and metrics such as precision-recall curves and ROC AUC are useful to identify model performances.

We further test the model using different combination of features to see which feature sets perform the best. First we use only the aggregated auto sales data, to predict the directionality of movement of SAARTOT index.

**Data**: *Sales data from Top 10 manufacturers for United States of America*

We consider automotive sales data sampled on a monthly basis, sourced from Bloomberg. The sales were aggregated for each manufacturer by summing up the the total car sales, light vehicle sales and truck sales. Missing values were treated with Forward Backward filling algorithm, and outliers were treated using univariate tests.

We observe that Random Forest model performed the best in terms of predictive performance metrics versus other machine learning techniques and the linear model. We describe the classification report of Random Forest Model in Table 3.

**Random Forest** 0.6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0.0 | 0.60 | 0.93 | 0.73 | 40 |
| 1.0 | 0.62 | 0.17 | 0.26 | 30 |
| Micro Avg | 0.60 | 0.60 | 0.60 | 70 |
| Macro Avg | 0.61 | 0.55 | 0.49 | 70 |
| Weighted Avg | 0.61 | 0.60 | 0.53 | 70 |
| Average Precision Score | 0.4613 | | | |

**TABLE 3. Classification Report for Random Forest**

ROC AUC:

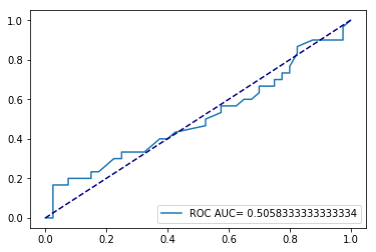


FIGURE 7. ROC AUC Plot of Random Forest

**Data:***Using the previous shifted absolute change*

The models do not perform much better than the original data by including the shifted absolute change. We observed that though the accuracy measure did not perform well , but ROC AUC improved by including the absolute change; the distribution of classification can be compared using classification of this experiment with the previous one.

**Random Forest:**

**Accuracy :** 0.557

**classification\_report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0.0 | 0.57 | 0.93 | 0.70 | 40 |
| 1.0 | 0.40 | 0.07 | 0.11 | 30 |
| Micro Avg | 0.56 | 0.56 | 0.56 | 70 |
| Macro Avg | 0.48 | 0.50 | 0.41 | 70 |
| Weighted Avg | 0.50 | 0.56 | 0.45 | 70 |
| Average Precision Score | 0.42667 | | | |

**TABLE 4. Classification Report for Random Forest Using Previous Shifted Absolute Change**

ROC AUC:

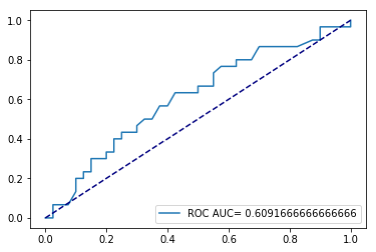


FIGURE 8. ROC AUC Plot of Using the Previous Shifted Absolute Change

**Data:** *Using Seasonally Unadjusted Data* We also experimented with seasonally unadjusted data, but it also did not improve the performance. In this case, ensemble models are performing worse than the linear model. One of the possible reasons can be that there are not sufficient features for ensemble models to perform well.

Performance of the Logistic Regression:

Accuracy: 0.5428

Classification report :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0.0 | 0.56 | 0.93 | 0.70 | 40 |
| 1.0 | 0.25 | 0.03 | 0.06 | 30 |
| Micro Avg | 0.54 | 0.54 | 0.54 | 70 |
| Macro Avg | 0.41 | 0.48 | 0.38 | 70 |
| Weighted Avg | 0.43 | 0.54 | 0.42 | 70 |
| Average Precision Score | 0.4266 | | | |

**TABLE 5. Classification Report for Random Forest Using Seasonally Unadjusted Data**

ROC AUC :

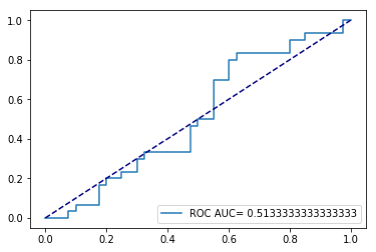


FIGURE 9. ROC AUC Plot of Logistic Regression using Seasonally Unadjusted Data

**Data:** *Using Additional Data*

Many other factors can also improve the performance, so we tried including more features which we thought may contribute to forecast better. To find out which factors are contributing most, we used backward elimination and forward selection feature selection techniques. We found that and contribute most while building ensemble models, and the least significant features are ,& index, index. We also found the distribution of for movement increase and movement decrease was different, the relationship was further explored in the bivariate section , a possible cause to find this feature most important. At that time, we also noticed that M1 INDEX, M2 INDEX has a lot of missing values, and missing values can be one of the reason that these factors are not improving the performance in any of the models we tried with. Usually, these index may have an impact towards the auto-sale, but the time reaction of auto company could be long. This can be another reason that they do have any contribution in forecasting the auto index.

While using the additional data, we did not use the features which are not improving the performance, and we found out that the new features helped improve the accuracy significantly for both the ensemble models, more than 7%. It indicates that addition data helped reduce both bias and variance. We observed that random forest performed better in terms of all performance measures than gradient boosting.

Random Forest:

Accuracy : 0.6857

Classification report :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** | **Support** |
| 0.0 | 0.68 | 0.85 | 0.76 | 40 |
| 1.0 | 0.70 | 0.47 | 0.56 | 30 |
| Micro Avg | 0.69 | 0.69 | 0.69 | 70 |
| Macro Avg | 0.69 | 0.66 | 0.66 | 70 |
| Weighted Avg | 0.69 | 0.69 | 0.67 | 70 |
| Average Precision Score | 0.555 | | | |

**TABLE 5. Classification Report for Random Forest Using Additional Data**

ROC AUC :

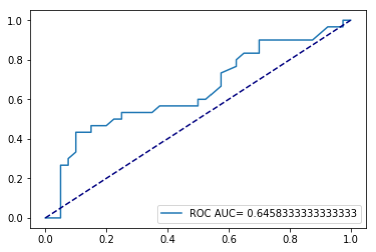


FIGURE 10. ROC AUC Plot of using Additional Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic Regression | SVM | Random Forest | Gradient boosting |
| Auto Sales Data | Accuracy : 0.5    Average Precision Score: 0.439  ROC AUC: 0.555 | Accuracy :0.5714    Average Precision Score: 0.428    ROC AUC: 0.5 | Accuracy : 0.6  Average Precision Score: 0.4613    ROC AUC: 0.5058 | Accuracy 0.542  Average Precision Score:  0.4232    ROC AUC: 0.5641 |
| Using the previous shifted absolute change | Accuracy: 0.5142    Average Precision Score: 0.443  ROC AUC: 0.5458 | Accuracy :0.5714    Average Precision Score: 0.428    ROC AUC: 0.5 | Accuracy 0.557    Average Precision Score: 0.4267    ROC AUC: 0.6091 | Accuracy 0.5714    Average Precision Score:  0.445    ROC AUC: 0.6216 |
| Using seasonal unadjusted data to predict[¶](http://localhost:8888/notebooks/Documents/yi/fin/FinancialModeling/code/Final.ipynb#Using--seasonal-Unadjusted-data-to-predict) | Accuracy: 0.5428  Average Precision Score: 0.4226  ROC AUC: 0.5133 | Accuracy :0.5714  Average Precision Score: 0.438  ROC AUC: 0.4725 | Accuracy 0.514  Average Precision Score: 0.4157  ROC AUC: 0.536 | Accuracy 0.528  Average Precision Score:  0.419  ROC AUC: 0.523 |
| Using Additional Data | Accuracy: 0.5428    Average Precision Score: 0.460  ROC AUC: 0.5616 | Accuracy :0.5714    Average Precision Score: 0.428  ROC AUC: 0.5 | Accuracy 0.6857    Average Precision Score: 0.555    ROC AUC: 0.6458 | Accuracy 0.6    Average Precision Score:0.4687  ROC AUC:0.61 |

Given sufficient number of features, we observe that ensemble models always perform better than linear models in our experiments in terms of metrics such as ROC AUC, Accuracy, Average Precision etc.

We also observe the Random Forest model slightly outperforms Gradient Boosting in most cases. This result might be due to the ability of Random Forest models to be robust to noise in data. Gradient Boosted Trees model is known to overfit the data, and might be overfitting on noise leading to a worse generalization error than the Random Forest Model.

# **CONCLUSION**

We observe in the experiment section that Machine Learning Techniques significantly outperform Linear Models in predicting the directionality of movement of Automotive indexes. With our experiments, we observe that Random Forest method performed the best amongst the techniques we evaluated on the basis of outsample performance. We also observe that features such as lagged sales of automobiles, GDP, Gasoline prices etc has the greatest impact on our predictions.

Though we collected the data from a Bloomberg, which is considered amongst the most reliable and clean sources of financial data, we observe that there were many missing values in the features. In general, to maximize the predictive performance of Machine Learning models we need the maximum amount of clean data, though a model fitted on a smaller but cleaner source of data would be typically found to perform better than a model fitted on a large but noisy data.

One of the future work in this research can be to extend the model to include additional data source and evaluate the impact, and another equivaly promising direction is to find a cleaner source of data.

Due to the overfitting nature of Gradient Boosted Trees (because it constructs trees on the basis of misclassified samples by its predecessor trees, and because the misclassified samples might be just random noise), Gradient Boosted Trees work better than Random Forests when the Data is clean, though Random Forest method is more robust to noise in data.

Another direction of work could be to extend the classification framework to regression framework, and evaluate the value add of predicting the exact value of the automotive index rather than the if the index is moving up or down.