

Project Hackathon Defense

Sales Forescasting for Retail clothing products

HAROON AHMED | BATCH 29 | 24th NOVEMBER,2017

# Purpose

The Aim of the project is to effectively forecast the sales of the clothing products for a leading retailer based on past sales history, to allow them to effectively place their product in stores. Sales forecasting is a widely-used technique for businesses to enhance their reach in the market by effectively allocating their resources. The project aims to deploy Machine learning and Time Series based models to perform sales forecasting and to do a comparative study on both the approaches.

Below are some of the areas the Business may benefit from the project: -

* Workflow Management
* Cashflow Management
* Resource allocation
* To provide discount offers to boost sales during off periods of sales
* To tap market opportunity
* Benchmarking against competitors

# Data

The data for this project comprises of the below mentioned elements.

* Sales (In Thousand Dollars) achieved in the last 7 years monthly for the product categories: Men Clothing, Women Clothing, Others Clothing. This dataset was utilized in developing both the Time Series and Regression models.
* Macroeconomic data for the past 7 years comprising of elements such as Combination of Year and month, Monthly Nominal GDP Index In Million Dollars, Monthly Real GDP Index In Million Dollars, CPI, Political party which is in power, Unemployment rate, Commercial Bank Interest Rate on Credit Card Plans, Finance Rate on Personal Loans at Commercial Banks\_24MonthLoan, Earnings or wages in dollars per hour, Expenses for ads in thousand dollars, Cotton Monthly Price US cents per Pound lbs., Percentage Change In Monthly Cotton Price Average upland Cotton planted In Million Acres, Average upland Cotton harvested In Million Acres, Cotton yield per harvested acre ( in pounds i.e.., lbs.), Cotton Production In 480\_lb net weight in Million Bales, Cotton Mill Use In 480\_lb net weight in Million Bales, Cotton Exports In480\_lb net weight in Million Bales
* Weather Data for the past 7 years comprising of elements such as Year, Month, Day, Temperature High(°C), Temperature average (°C), Temperature Low (°C), Dew

Point High (°C), Dew Point Average (°C), Dew Point Low (°C), Humidity High Percent (°C), Humidity Average Percent (°C), Humidity Low Percent (°C), Sea Level Pressure High(hPa), Sea Level Pressure Average(hPa), Sea Level Pressure (hPa), Visibility High(km), Visibility Average(km), Visibility Low(km), Wind Low(km/h), Wind Average(km/h), Wind High(km/h), Precipitation sum (mm), Details of weather like snow, rain, fog etc.

* Events and Holidays for the past 7 years comprising of features such as Year, Month and date combination, Details of special event or holiday, Whether federal holiday or event.

## Data Challenges

The major challenge for the Regression problem was posed by the data being available as separate excels. A considerable amount of time was thus spent to combine the data by aggregating and processing them to single format based on year and months.

The Weather data had lot of features with many of them being debatable in terms of their relevance to Women shopping patterns and subsequent sales. Feature selection and engineering was a key challenge in the weather dataset provided.

The Macroeconomic data had lot of features that seemed similar and posed a challenge of multicollinearity for the problem. Features such as Monthly Nominal GDP index and Index and Monthly Real Nominal GDP index were prime example of these. Unemployment and Monthly Nominal GDP Index on the other hand showed strong negative correlation among them. An attempt to fix these was made using a custom function which deployed the VIF (Variance Inflation Factor) but its potential couldn’t be exploited in the interest of time.

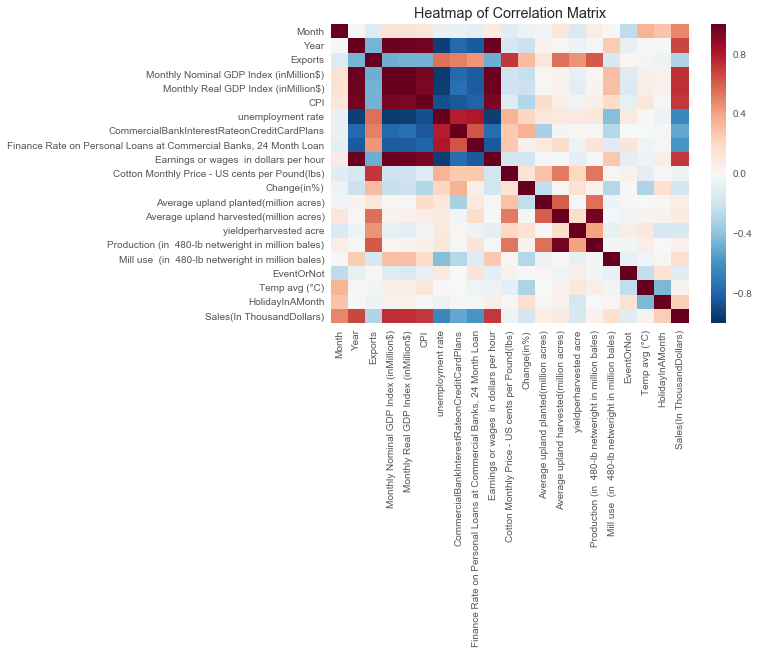


Fig. Correlation Matrix indicating strong correlation among few features

The Events and Holiday data on the onset tended to provide the key features for Sales. As from intuition it can be determined that the inclusion of Events and holidays feature in the train dataset would be key to developing an accurate forecasting. Thus, considerable amount of time had to be spent to feature engineering elements from this dataset.

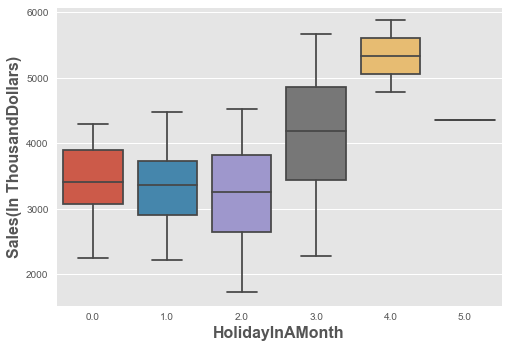


Fig. Holiday in A Month feature in response to Sales

## Additional Data

Data such as Months during which discounts were offers provided by the retail store, Hourly visiting hours of customer, detailed Advertising expenses could be further added to enhance the accuracy of the model.

## Preprocessing

##### Regression

* Dropping Features: Many features were dropped based on similarity to other features, difficulty in processing or irrelevance. From Weather data, Average Temperature, Event Ratio was chosen to be taken forward for processing rest of the data was removed. In Events and Holiday data, Details of special event or holiday was removed citing the difficulty it possessed to process them. In Macroeconomic data, **PoliticalPartyInPower** feature was dropped as it possessed a single distinct value and had no variance which could be learnt by the Machine learning model. Advertising expense was also dropped as it had more than 40% of the data missing.
* Data Imputation: Simple mean was to impute values wherever necessary
* Standardization: The Final dataset was standardized to scale all the values between 0 and 1 as some of the features like Monthly Nominal GDP Index In Million Dollars had high values compared to other features. But it was observed during prediction on Test data set that standardization led to significant dip in accuracy thus final models were developed without standardization.
* Dummification: DayCategory feature was the only categorical feature in the data which was dummified to be utilized in the Machine learning process.

##### Time Series

* Data Imputation: Central Imputation was used to impute values wherever necessary

## Feature Engineering

##### Regression

Some of the additional Features generated from the provided datasets are as below:

* Holidays in a month: Obtained by dummifying and aggregating the DayCategory feature based on month in Event-Holiday dataset to obtain the number of holidays in a month
* Weather Event Ratio: Obtained from the Weather event by aggregating based on months and by converting empty field as 0 and rest of the field based on string length. The feature indicates a high ratio oscillating between 0 and 1 if the frequency of rain, fog and snow are high in a specific month.
* Month and Year extracted from Year-Month Feature in Macroeconomic Data. Month is then converted to numerical to simplify of merge with rest of the data.

##### Time Series

* Data was available in ready to use format with feature engineering unnecessary.

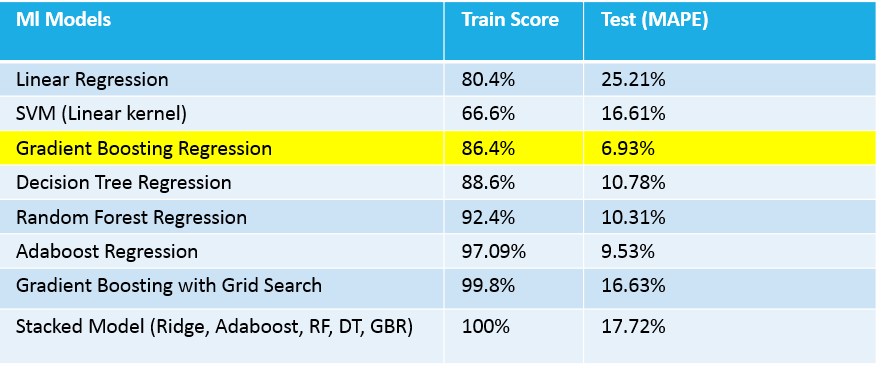
# Approach

## Regression

The aim of the problem was to accurately forecast the sales of clothing products for women for the year 2016. The approach followed to solve the problem was to aggregate all the relevant features from Weather, macroeconomic and train dataset into a single dataset to ease model building on top of it. Standardization, Train-test-Validation split, imputation and Dummification were some of the preprocessing as mentioned above that were used in the process.

##### Model Building Experimentation

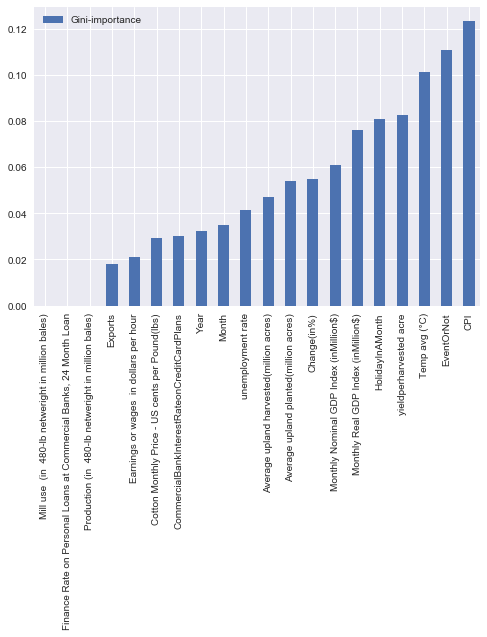
The below models were deployed for regression on the processed dataset with the resulting train accuracy recorded as below:



The best accuracy on test data was thus obtained on Gradient boosting regression method.

##### Bottleneck Analysis

Most of the models build on the Final dataset being of dimensions 84 rows and 20 columns overfitted when complex models were used and with Grid Search of parameter space as is evident from the train score. Stacked model could achieve 100% score on train dataset exemplifying the issue. The simpler models failed to generalize well on unseen data. Gradient boosting Regression was thus used as the final model deployed as it did not overfit on train data and generalized well over unseen data. One of the reason being its ability to give importance to multiple features rather than basing its decision on limited number of features.



**Fig. Feature Importance – Gradient Boosting Regression**

## Time Series

The aim of the model was to accurately forecast the sales of clothing product for Women, Men and Other category for the year 2016. The approach to building the model for this problem was to handle the three categories separately by building Time series model for each category and combining the predictions for testing on unseen data. The data was observed to have strong seasonal patterns and where additive in nature.

##### Model Building Experimentation

The data had missing values. Central Imputation and KNN imputation were chosen for experimentation to impute the missing values. The central imputation method was used in model building based on higher accuracy on unseen data. ARIMA and Exponential Smoothing models were used as Time series models on the train data. With both giving equal accuracy on the test data. Thus, ARIMA model was chosen to the final model.

##### Bottleneck Analysis

Time Series of Other Category had significant spike in its time series data which could be standardized to obtain a more generalized model.

# Results and Analysis

The results obtained from Time Series and Machine Learning model when compared suggests that better results and explainability of Data can be achieved much more effectively from the causal variables used in the Machine learning model, as evident from the table below:

|  |  |
| --- | --- |
| Model | Test Accuracy |
| ARIMA | 8.45% |
| Gradient Boosting Regression | 6.93% |

##### Shortcomings

* Train-Test-Validation split couldn’t be carried out effectively as the model took a hit in terms of accuracy when some data was retained for validation. Thus, Train accuracy and test error needed to be relied upon for accuracy estimation

##### Improvements

* Feature selection using methods like StepAIC, VIF or the important features generated from Tree based models could have been carried out better given more time.
* More regularized could have been built given more time.

##### Lessons Learnt

* Data preprocessing is key for any model to work well and enough time should be allocated towards it.
* Need to be careful about overfitting when dealing with complex models

# Appendices

Find attached below the Python Script for the regression model developed and the R script for the Time Series model developed.



