TITLE: PREDICTING WALMART SALES

AUTHORS:
1. Nikhil Junneti, SBU ID: 110750943
2. Shravya Rani Thatipally, SBU ID: 110739889
3. Gopi Krishna, SBU ID: 110739515
4. Rajendra Vellisettti, SBU ID: 110479075

MOTIVATION AND BACKGROUND:

Sales forecasting is an important tool for any growing business especially to determine future revenue and planning for any demand. Walmart being an American multinational retail corporation, the sales prediction for it will have huge impact on many different peripheral departments, and will help make many important business decisions. Accurate sales forecasting can help you track data to gain insight into areas where improvements can be made.

Knowing where revenue has been historically generated and where new sales are coming from will help an organization make statistically sound judgments regarding future performance. These results can be benchmarked to forecast revenue and demand. This will allow an organization to plan operations better and have clearer insight into your overall supply chain and deliverables.It can adjust your operations to run more in line with those learnings

METHOD:

Data exploration: We have taken sales data for 45 Walmart stores located in different regions. Each store has number of departments. We have information about weekly temperature, fuel price, CPI, Unemployment rate in the store region, offers running in the store, and whether week has a special holiday for each store. Also, our training dataset has weekly sales information for each department. Our goal in this project is to predict department-wise weekly sales for each store.

Data Processing: Initially we have merged the available training data and features to focus on department sales. All markdown data(there are four columns containing sales data on holidays with special offers) available is summed up and taken as offers. We added new column called WeekNo which indicates the week no corresponding to that year(Varies from 1 to 53). The data in the holiday column is converted from boolean value to binary value to adapt to regression models. We then standardized the weekly sales data for each store using corresponding store's mean and standard deviation.

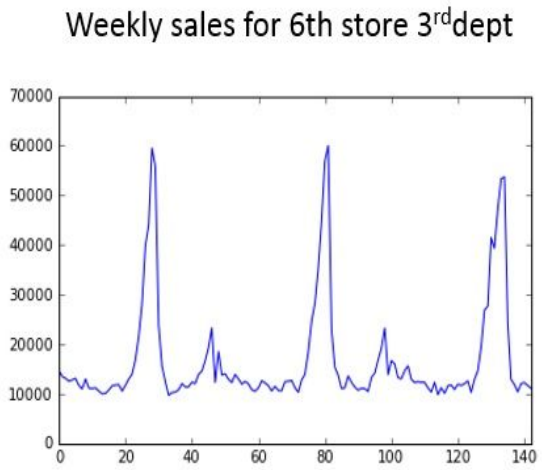
We calculated the correlation between sales and all available features and found that none of them are significantly correlated with sales as shown in the following correlation matrix. This fact is also verified by hypothesis testing.

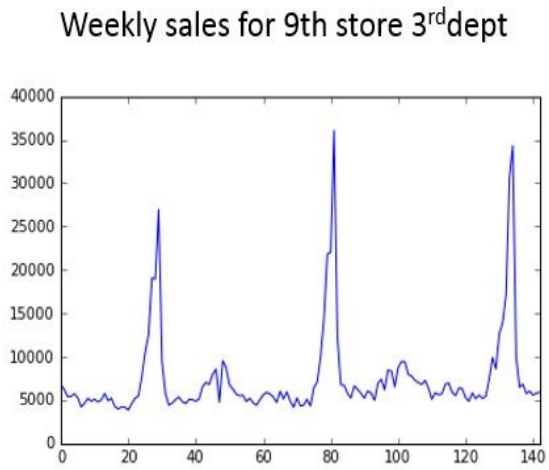
Correlation Matrix

	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	CPI	Unemployment	offers
Weekly_Sales	1.000000	0.059404	-0.372496	-0.004577	0.055249	-0.023030	0.005297
IsHoliday	0.059404	1.000000	-0.183660	-0.085903	-0.028945	0.046214	0.460340
Temperature	-0.372496	-0.183660	1.000000	0.267038	0.130900	-0.106509	-0.236168
Fuel_Price	-0.004577	-0.085903	0.267038	1.000000	0.754206	-0.384444	-0.263664
CPI	0.055249	-0.028945	0.130900	0.754206	1.000000	-0.831687	-0.332047
Unemployment	-0.023030	0.046214	-0.106509	-0.384444	-0.831687	1.000000	0.282780
offers	0.005297	0.460340	-0.236168	-0.263664	-0.332047	0.282780	1.000000

Time Series Analysis: We then tried to predict the sales value by observing the trend over time. We found that similar patterns are repeated every year and also these patterns for a department is similar across all stores. Below is the sales value plot for two different stores for same department.

SALES PLOTS





With the above plots, we inferred that previous years' pattern has a huge impact on current year's prediction for corresponding week sales. And also same trend is observed across all stores for a department. To capture the trend over all stores we have taken average sales value for all stores. But each department sales are in different range, so we have standardized sales within each department.

To capture seasonal trend for a store we have taken average sales value for the same week number in previous years

To capture the trend for all stores we have taken average sales value over all stores for the same week number in the previous years.

The previous week sales is taken after grouping the data by store and is then standardized

So based on these plots and observations, we considered the following features:
avg_overall_week_sales,
avg_overall_next_week_sales,
avg_overall_prev_week_sales ,
avg_store_week_sales ,
avg_store_next_week_sales,
avg_store_prev_week_sales for all previous years and
previous_week_sales for current year.

Hypothesis testing for significant features: We have applied regression on these features and calculated beta values. We then performed P-test on these values and found them to be significant.

Model selection and prediction: We have constructed model for each department. In the model selection we have considered Lasso, Ridge and OLS regression models. We have used 10-fold cross validation for testing the models. Amongst these models OLS is well fitted for our data.

From the final model we have calculated deviations of predicted values from the actual sales (error). Then we tried to model this error based on the environment features of stores.

PROJECT CRITERIA:

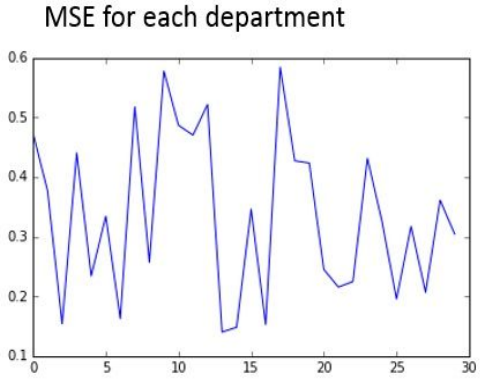
1. **Probability Theory:** We calculated the probability of increasing or decreasing sales by at least 25% wrt previous week sales. We calculated these probabilities given the week has special holiday. We found that information about week having special holiday is not effecting the probabilities much.

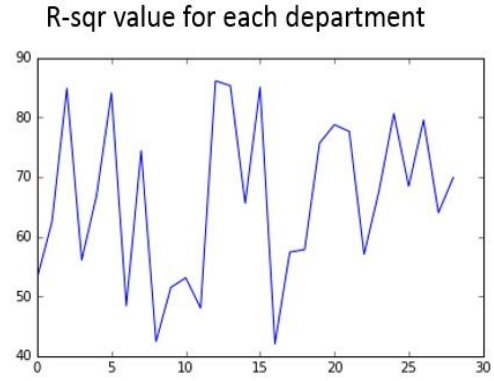
2. **Discovery:** We calculated correlation matrix of each feature on sales. We then performed significant test on all the beta values. Based on the p-values obtained from the test, we figured out that available features are not sufficiently significant features. So, extracted new features which have sufficient p-values that actually affected the sales value.
3. **Prediction or Clustering:** We compared multivariate linear regression, lasso and ridge regression by applying 10-fold cross validation on each of the models and calculated MSEs. We finally selected OLS model whose MSE is the lowest. OLS model fits best to

our sales prediction. We also have used this to model the error from environment features.

EVALUATION AND RESULTS:

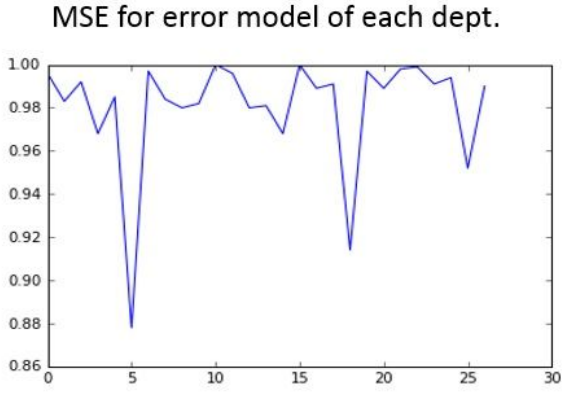
Evaluation metrics: MSEs and R-squared values were considered for evaluating the accuracy of model.

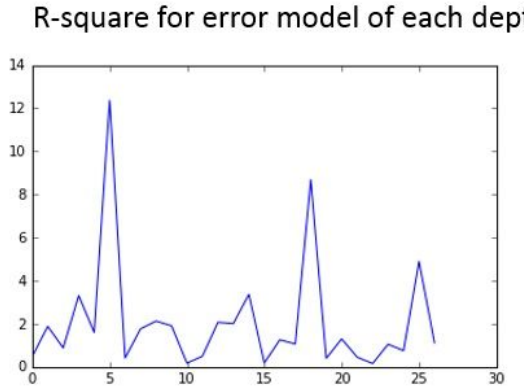




MSE and R-Square values for each departments are shown above. This plot shows that most of the department models MSEs are below 0.5 showing that in general our features works well for all departments. Also, the R-square values lie in the range 60-90. This is another indication that supports our prediction.

Also, we tried to predict the residual error using the environment variables which are given along with data (temperature, fuel price etc). We tried to explore this correlation in various departments and found that these features are almost insignificant in predicting the error for most of the departments.





CONCLUSION:

a) Effect of given features is not significant. Time Series analysis played a vital role in predicting sales.
b) An accuracy of 0.1-0.5 MSE and 60-90 R-square value for most of the departments shows that those departments in particular and all departments in general should estimate the demand and supply based on some important factors like previous week sales for current year, avg_overall_week_sales, avg_overall_next_week_sales, avg_overall_prev_week_sales, avg_store_week_sales, avg_store_next_week_sales, avg_store_prev_week_sales for previous years.
c) Sales in and around holiday weeks are exorbitantly high showing the same pattern over years. From this, it is clear the supply and demand can be well planned for those weeks in every year. Sales in some departments of walmart stores rely on environment variables as predicted from residual error.

With all the above observations, we believe that sales prediction can take into account our results for improving their planning for supply and demand.