**Big Mart Sales Prediction**

**A Project Report**

***Submitted by***

**Gangadasu Vinoothna – 1700209C203**

**Nimasa Konsam – 1700249C203**

**Charishma Pydi – 1700203C203**

**Batch - 2017 CSE 2**

***Course***

**MACHINE LEARNING and DATA MINING - CSE2702**



**SCHOOL OF ENGINEERING &TECHNOLOGY**

**BML MUNJAL UNIVERSITY**

**April 2020**

**Acknowledgement**

The project that we choose consumed huge amount of work, research and we had a great chance for learning and implementation. In order to make it success we got a lot of support from Dr. Atul Mishra and we are very thankful to you sir for taking part in useful decision and giving necessary advices and guidance and arranged all facilities to make life easier during our period. We are so glad to use this opportunity to convey our wholehearted gratitude. We recognize this opportunity as a big milestone in our future development. We will strive to us acquire skills and knowledge in the best possible way, and will continue to work on their improvement, in order to attain desired career objectives.

**Abstract**

The companies which are engaged in manufacturing, wholesaling, sales forecasting is a prime aspect. Because sales forecasting gives the analysis of company which helps in future growth of the company, it estimates the sales, plan the strategy for the growth. And the sales forecasting also gives you the knowledge about the market and market place. So in simple words we can say that machine learning algorithms which helps us to predict the values based on the past data and give the predictions for upcoming years which leads to the company growth. The main goal is finding the sales of each and every product on ten different stores and to build a model which is very predictive. So basically the steps that were involved are Hypothesis, Exploration of Data, Data cleaning or Data pre-processing, Feature engineering and applying the linear regression, decision tree based algorithms. And finding the accuracy levels of each model that we used and also doing some changes to give more accurate results.

**Keywords**

Data Exploration, Data Pre-processing, Feature Engineering, Hypothesis, Machine Learning, prediction, Sales forecasting.

**Introduction**

Sales are most important for one’s business. A sale forecasting which also plays a major role for conducting any kind of business. If that forecasting gives you an accurate value or result that will helps you in business growth if the prediction is false then it will effects the business. And the sales forecasting also gives you the knowledge about the market and market place and also estimates the marketing plans for future. Basically sales forecasting comes under predicting the sales for upcoming years. Businesses or companies who are following the sales forecasting or sales prediction or taking these seriously are having the good growth than the people who are not.

In our project we started with brainstorming the ideas like what actually related to sales prediction. Display area, location, capacity of the store, stock of the products, stores on particular area is all comes under hypothesis. We have to focus on this hypothesis for example we can see in daily life that many of us don’t want to stop on multi stores to buy things because of many reasons and the people wants to buy all the necessary goods from only one store if available. So this comes under store capacity. We applied the pre-process on big mart data which consists of unique 1559 products at ten different stores by filling missing values with mean values. And then we apply feature engineering which means making some changes in the existing values and then analyse. After that applying the models to that existing ones the models which includes linear regression, decision tree. And getting the accurate results. The performance metric that we used to check our models is root mean squared errors.

**Literature Review**

Big Mart Sales Prediction in which we are predicting the sales which is discussed in [1]. As we all known that sales market plays a major role in developing countries which is discussed in [3]. Sales are most important for one’s business. A sale forecasting which also plays a major role for conducting any kind of business [1]. If that forecasting gives you an accurate value or result that will helps you in business growth if the prediction is false then it will effects the business. And the sales forecasting also gives you the knowledge about the market and market place and also estimates the marketing plans for future which is discussed in [1].

The dataset that we collected consists of 1559 unique products on ten different stores. And we are following the following sequence in order to get accurate results. Hypothesis->Data exploration->Data pre-processing->Feature Engineering -> Models (linear regression, Decision tree based) which was discussed in [1]. In order to get accurate results we have to follow these five steps in a sequence.

After the hypothesis is made we have to explore the data by the data exploration [1] and we have to compare the original data with the hypothesis (75% of attributes are matching). By the data exploration we have to check whether we can modify any values or create any attributes in order to get accurate results which are discussed in [1]. Feature Engineering, in this step we actually modify the values or attributes that we declared in data exploration step [1] [3]. In data exploration step we just gave a hypothesis. In this step it is actually turned into reality.

Finally we reached to model stage in this stage we have to apply the models and check the accuracy which is discussed in [1]. Coming to linear regression, we have to find the r-value [2] of the dataset which says that how much of dependent variable depends on independent variable [2] [1] this is very simple and effective. Decision tree based algorithm is a supervised learning [2] based on non-parametric [2] that is used for regression [1] and classification [1] tasks. For this project we used these two algorithms. Training dataset is used to train the classifier and then test that classifier with the help of test dataset and find the accuracy. The accuracy is totally depends on the amount of the data that is provided to the classifier and the attributes selected which is discussed in [1] [3].

Two level statistics [1] we can apply the algorithms like linear regression, ridge regression, decision tree regress or which gives predicted value that takes an input to the top layer which we can say second layer and gives the final prediction based on the input that we took from the bottom layer [1]. Top layer we are using the models like cuboid [1]. Companies who are using these two level statistics are getting more benefits compared with the single level layer [1].

In [4], about four sub-models were used whose output was used as an input into two variants of the neural network(single and multilayer) to create a hybrid model. The sub-models were Auto Regressive Integrated Moving Average(ARIMA), Simple Moving Average(SMA), Support Vector Machine, Radial Basis Function. It showed how useful the complex models can by creating a hybrid of several algorithms.

But the usage of algorithms totally depends on the type of problem we solve. In [5]The black Friday sales were predicted using a predictive model was built with linear regression, logistic regression, Stepwise regression, lasso regression, ridge regression, elasticnet regression and could achieve high accuracy. The data after collection was processed thoroughly by filling missing values and one-hot encoding. It showed how complex models like neural network fail simple problems. Simple problems with proper data cleaning can work miracles with algorithms like regression.

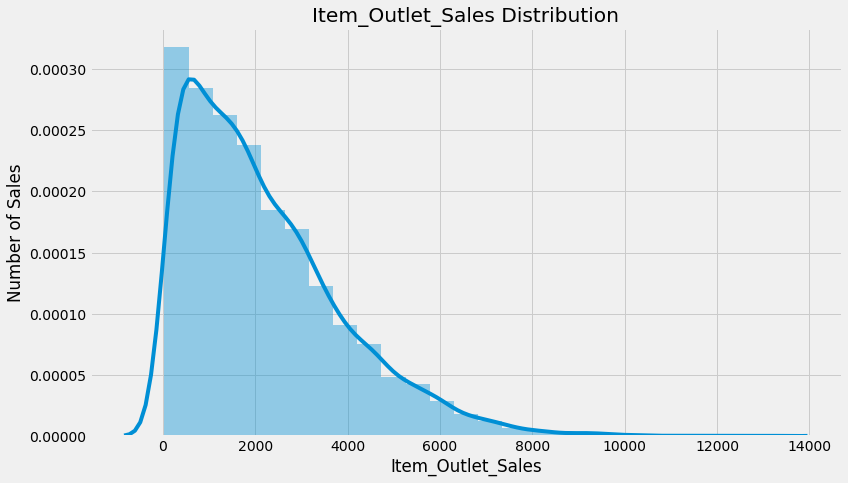
Any business organization uses such prediction to model its sales accordingly. In [6],[8] previous years data was collected, processed and then performed validation and implementation using accuracy models like MAE and RMSE. This helps in the prediction of the following years sales and helps in modelling their business according to what the customers might want.

In any case, for a machine learning algorithm to perform well, data is the building block. Data mining plays an important role in forming the knowledge base for businesses. In [7] the accuracy and performance issues of the machine learning algorithms were overcome by using various data mining techniques. It involved gathering previous years data and then exploratory analysis. After data understanding, preparation, modelling, evaluation and deployment, outlier detection is done.Later predictive models like linear regression, decision trees, gradient boosted trees were used. But it is undeniable that business decisions depend on the speed and accuracy of data processing and data visualization techniques.

**Methodology**

In Big Mart Sales which has a dataset. The dataset which consists of unique 1559 products at ten different stores we are following the five steps Hypothesis, data exploration, data pre-processing, Feature engineering and models.

From the below graph we can see that Outlet\_sales is skewed to right. The main goal is to change the curve to symmetrical we could change this by using square root.



**Hypothesis:**

Firstly, Hypothesis in this step we mainly focus on brainstorming as we mentioned earlier in introduction that impacts the result. This step is focuses on brainstorming and understanding the problem before considering the actual data and came up with list of possibilities that we have to consider. So here we are showing the list of hypothesis that we made. So when we think of sales the general thinks that comes to mind are products and stores. So we made two level hypothesis that is product and store.

**Store level Hypotheses**

**City type**: Compared to rural areas the cities which are developed have higher sales because of higher population and because of rich people.

**Population density:** Stores which are in cities and highly populated areas have higher income compared with lowly populated areas.

**Store capacity:** As we already mentioned this example above that people prefer the stores which are large in size.

**Competitors:** If we placed our store in an area which consists of more than one same type of stores then that leads to lower sales because of the competition.

**Location:** Location is the most important aspect that we have to consider and we have to place our store where the demand should be high.

**Ambiance:** most of the people among us give importance to the ambiance than the products in the store. So we have to behave in a polite manner with our audience

**Marketing:** good marketing leads to good sales. Because we have to attract them by amazing and right offers.

**Customer Behaviour:** The way of receiving the customers is also most important. It shows impact on sales.

**Product level Hypothesis:**

**Brand:** Customers trust the branded products. So we can say that it also has an impact on the sales.

**Packaging:** Customers could be attracted by good packaging which leads to higher sales.

**Display area**: Bigger shelves are likely to catch attention. so the display area should be good.

**Utility**: Demand for daily utility things should be more compared with less utility things.

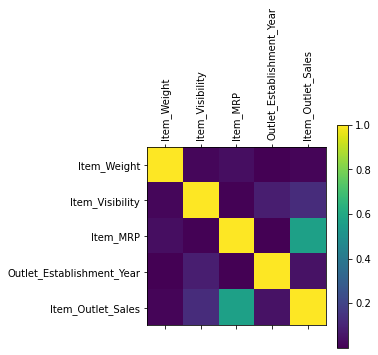
**Visibility in store**: In many stores we saw that if there are any offers or any new products they would like to place it in front of the stores to catch more attention and they increase their sales by this method.

**Advertising**: Advertising is most important in stores as well as in products factor also.

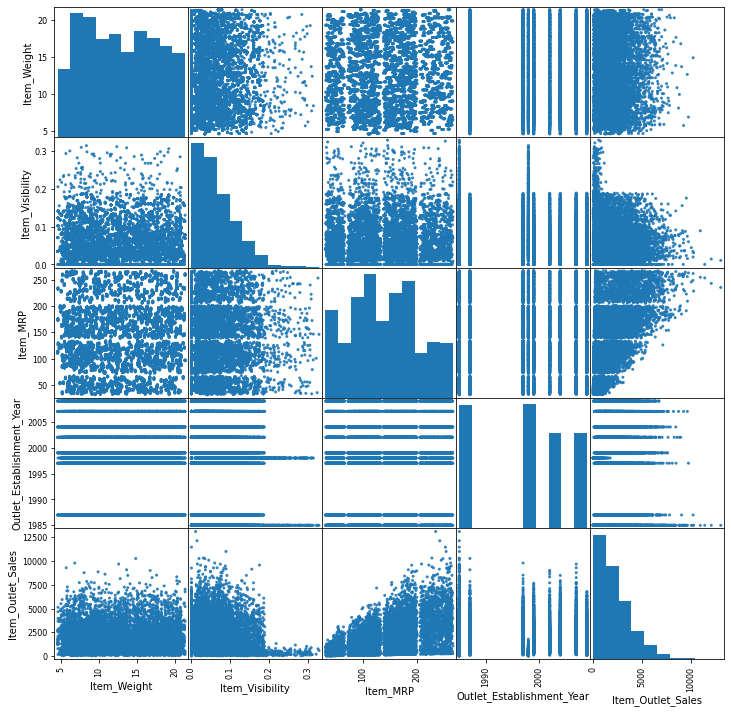
**Promotional offers:** If we provide the offers to products that products will be sold more compared to other products.

**Data exploring:**

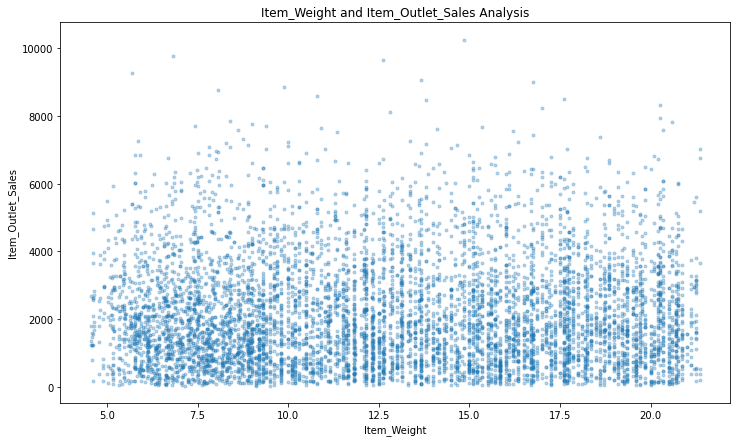
After the hypothesis is made we have to explore the data by the data exploration By the data exploration we have to check whether we can modify any values or create any attributes in order to get accurate results.

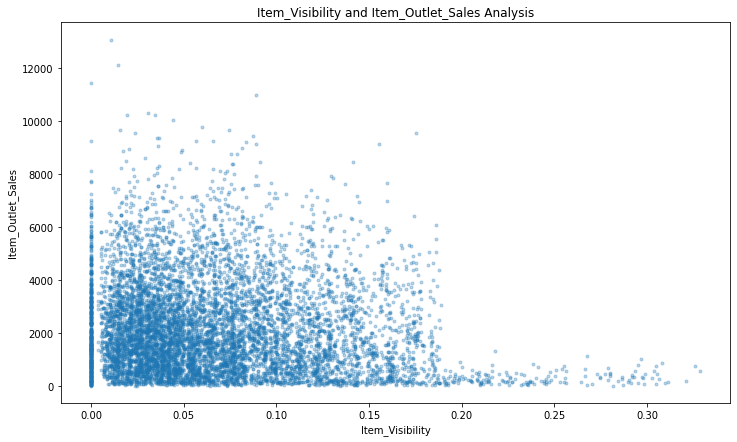


As we can see from this figure correlation between each attributes.

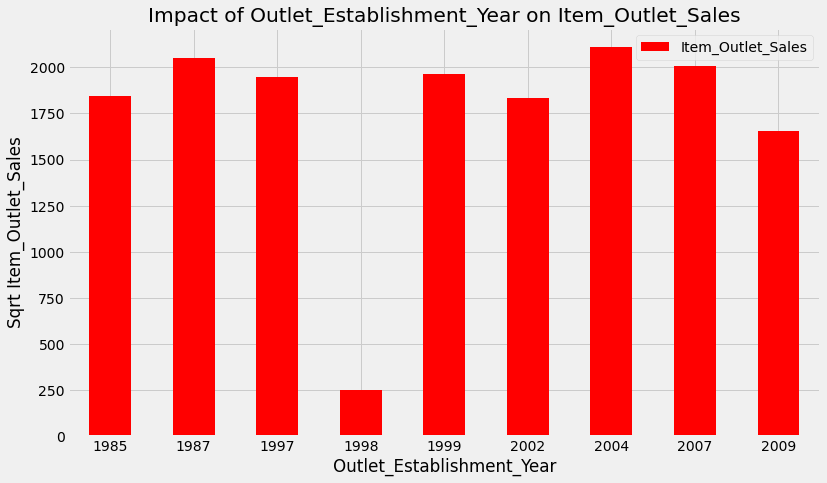


From the above figure we can say that there is no correlation between Item\_weight (independent variable) and Item\_outlet\_sales(dependent variables). If you want to see it in more clearly we can only the plot the graph between those variables.

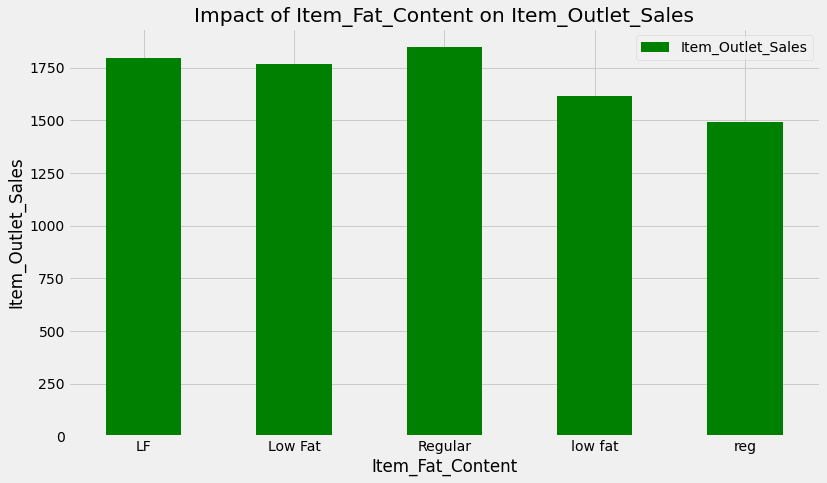
  
Item\_visibility in which we hypothesised that if the visibility is more, than the sales will be more. But in the graph below which we can say that if the visibility more, than the sales will be less. So from this we can say for the daily utility products there is no need of visibility



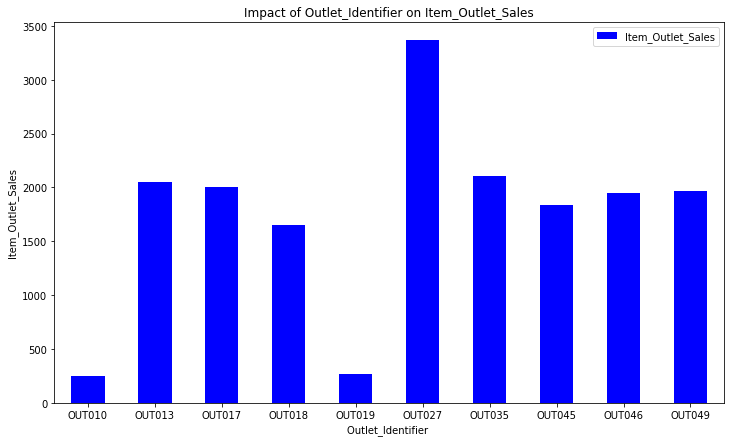
Outlet\_Establishment\_year vs Item\_outlet\_Sales there is no need of correlation between these two. In 1998 the sales are very low one possible reason would be the competition increases and the same type of stores is placed on that area.



Item\_fat\_content from the below graph we can say that the items which are having low fat has higher sales compared with the regular fat items.

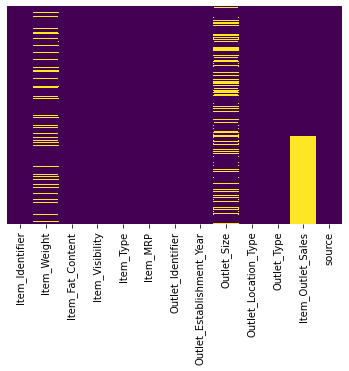


And we found which outlet\_identifier has higher sales. Out027 has higher sales compared to others.



Data pre-processing, coming to this step it plays an major role among all the steps because in data pre-processing which is also called as Data cleaning has to impute the missing values by some methods. We are following the most efficient which we impute the missing values with mean value of that attribute. The dataset that we downloaded is actually splits into training and testing dataset.

In order to make it fast we have to combine these two datasets into one which is called data dataset. The advantage of doing this is instead of applying the same methods for these two datasets separately we can do that better if it is combined. By the concat method we can combine these two datasets and we added another column which is called source. And the next step that we have to do is find out how many columns have missing values and then we have to describe the data by the following methods and their results are shown below.



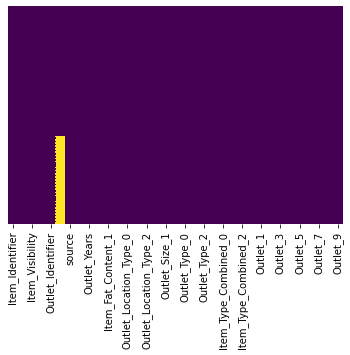
From this we can say that we have to impute the missing values in variables called “Item\_weight” and “Outlet\_size” and we ignored the “Item\_outlet\_sales” because in the test dataset that column will have empty values.

From this describe() method we found out count, mean, std etc.. but it says that Item\_Visibility is 0.0000 but practically it should not be zero. Item\_visibility comes under Display area that we hypothesized so if display area is 0 which means there is are no products which results in lower sales. Outet\_Establishment\_year can also be changed to values not like 1984, 2007 but it should be like 4,5,28 soon.

In this step we are finding the unique value of each and every variable. From this step we can say that the Item\_identifier has 1559 unique products and outlet\_Identifier identifies 10 different stores which exactly match with our problem statement. And one more observation source the column that we added and for testing dataset we gave it as test value and for training dataset we gave it as train value so it has 2 unique values.

**Data cleaning:**

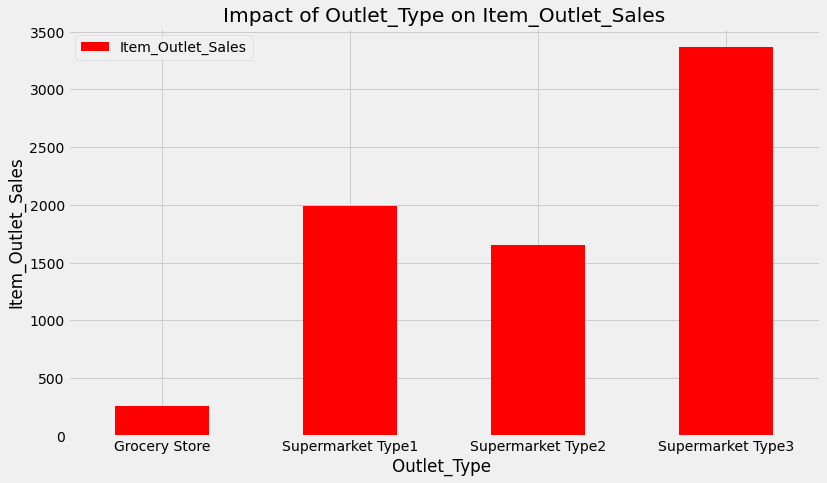
In this we mainly focus on imputing the missing values. As we already know that how many attributes contains the missing values. We will fill those missing values by the mean of that attribute. Item\_weight where we can impute the values by the mean but in Outlet\_size we impute the values based on the mode like small, medium, large. We categorized the size based on that mode. After importing the necessary values and modes we are done with imputing the missing values. As shown in the graph below.



**Feature engineering:**

After imputing the values we are done with the data pre-processing step. Here comes the feature engineering as we can say that feature engineering refers to modifying the existing values and analyse it by creating the new one.

1. We want to check whether it would be a good idea if we combine the supermarket type1 and 2. As we seen below that it shows completely different sales then it is not good idea to combine them.



1. As we know that visibility cannot be zero so we have to modify it and impute the values wherever there is a zero. And fill it with the mean values. In the hypothesis step we stated that higher visibility leads to higher sales. Along with that we have to look at the product in our store as compared to the average visibility of the product in all the stores. In order to achieve this we can use the variable called visisbility\_avg. For that we created a new attribute called Item\_Visibility\_Meanratio.
2. In Item\_type we are having different types of the item and in Item\_identifier every first two characters describes whether it is foods or drinks or non consumable products. We categorized this based on that.
3. Instead of determining it in years it is better to have it in min and max range of how old the store with respect to particular year that particular year as we already mentioned in problem statement.
4. In Item\_Fat\_content we already seen there is a kind of disturbance that we have Low fat, low fat, lf, which is nothing but same. So we have to take care of this now. And in this step for some items the fat content was not mentioned. On those items it comes under either low fat or regular category. So mainly for those items we created an attribute called Non\_edible
5. This is the important step as we know that sklearn only takes the numerical values but not the categorical values. So we have to convert all those to numerical values. So we converted in a way such that each variables in an attribute are stated as attribute suppose in Item\_fat\_content which is an attribute which consists of three variables that three variables are converted as Item\_Fat\_Content\_1, Item\_Fat\_Content\_2, Item\_Fat\_Content\_3. And the values that are present on that particular attribute can be stated as 1 else it will be 0. And one more point that we copied the oulet\_identifier to outlet and we made changes in outlet but not in outlet\_identifiers.
6. Now the data is ready in this step we mainly focus on exporting the data as we mentioned earlier we combined the two datasets to single dataset which is named as data dataset. And did necessary changes and now again we are splitting the data and then we exported the modified datasets.

**Models**

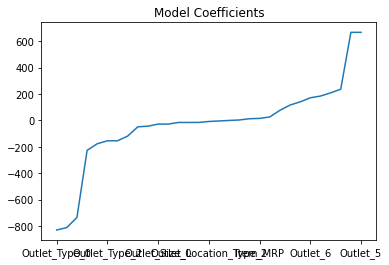
Coming to the last step as we stated that we are applying machine learning algorithms like linear regression, Ridge regression and Decision tree based algorithm. So firstly we started with applying the base line model it is nothing but calculating the overall average sales by product or in particular store by product. But it is not so efficient method so we applied the different algorithms in order to get the accurate results.

Regression is something used to find the good relationship between dependent and independent variable. Linear regression is something that is based on supervised learning and performs the regression task. Cross validation is something which is used to check the efficiency of the models and helps to tell whether our model is overfitting or underfitting or perfectly fitting.

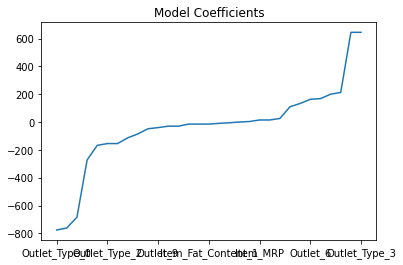
Based on the output we say that our model is overfitting so we applied Ridge regression. Ridge regression helps to tune the parameters, model coefficients and avoid or prevent the overfitting. And finally we applied the decision tree based algorithm. Decision tree based algorithm is a supervised learning based on non-parametric that are used for regression and classification tasks.

**Results**

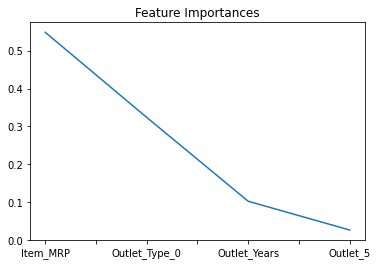
We got the modified datasets and then we applied the baseline model which is nothing but the mean\_Sales and it is not so efficient method. And then we applied the linear regression model which gives good results when we compared with baseline model. And the accuracy that it gives is 56.35.



We can see the overfitting of the model so we applied ridge regression which prevents the overfitting as well as tune the parameters or model coefficients. That we discussed above.



But the accuracy levels is 50 percent so we applied the another model called Decision tree based model where we get the accuracy levels upto 60 percent which is more better than linear regression model. And we can also see the model that actually fits into data. And in this model we actually consider the top four features where the target attribute (Item\_Outlet\_sales) depends on. The features that we considered can be seen in below graph. The accuracy levels is also increased because of that. We can also consider the other features but the accuracy differs. So based on the accuracy we only considered the top four features.



Below is the table of comparison of accuracy and cross validation root mean square of different algorithms that we applied in this project. Lower the root mean square error proportional to lower error this will increase the efficiency of model. So finally based on decision tree based algorithm we are predicting the sales of each and every product on ten different stores.

|  |  |  |
| --- | --- | --- |
| **Model** | **Cross Validation RMSE** | **Accuracy** |
| **Linear Regression** | 1127 | 56.75 |
| **Ridge Regression** | 1129 | 56.75 |
| **Decision tree based** | 1071 | 60.78 |

**Conclusion**

Firstly we made hypothesis then we explore the data and check the possibilities to modify the data then we pre-process the data by impute the missing values by mean then we apply the feature engineering where we check whether the modified data will give you good results or not if yes then only we proceed to modification. And then finally we applied different models to increase the accuracy of our predictions.

The method that we are following gave the accuracy levels upto 60 percent which we can say it is more efficient. In order to increase our efficiency we can also use random forest method, xGboost model which increases the accuracy of the result. The limitations of our project are single level prediction. We can extend this by two level statistical approaches. The other name of this two level approach is stacking where in the bottom layer we can apply more than one learning algorithms based on the predictions that we made it shows the accuracy and the top layer take those predictions as input and apply the learning algorithm called as cuboids and gives you the final prediction by taking the previous predictions as input which is more accurate than we compared with single level predictions.

**References:**

[1] Kumari Punam, Rajendra Pamula, Praphula Kumar Jain ,” A Two-Level Statistical Model for Big Mart Sales Prediction”, 2018 International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018

[2] Kunal Pahwa, Neha Agarwal ,” Stock Market Analysis using Supervised Machine Learning “,2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019

[3] ASHISH SHARMA, DINESH BHURIYA, UPENDRA SINGH, “Survey of Stock Market Prediction Using Machine Learning Approach”, International Conference on Electronics, Communication and Aerospace Technology ICECA 2017

[4] Usmani, M., Ebrahim, M., Adil, S. H., & Raza, K., “Predicting Market Performance with Hybrid Model”, 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 2018.

[5] [Ching-Seh Mike Wu,](https://ieeexplore.ieee.org/author/37086702517)[Pratik Patil](https://ieeexplore.ieee.org/author/37086700689), [Saravana Gunaseelan](https://ieeexplore.ieee.org/author/37086700717), “Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data ”, [2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)](https://ieeexplore.ieee.org/xpl/conhome/8649692/proceeding), 2018.

[6] T, G., Choudhary, R., & Prasad, S., “Prediction of Sales Value in Online shopping using Linear Regression”, 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

[7] Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S., “Intelligent Sales Prediction Using Machine Learning Techniques”, 2018 International Conference on Computing, Electronics & Communications Engineering, 2018.

[8] Gurnani, M., Korke, Y., Shah, P., Udmale, S., Sambhe, V., & Bhirud, S., “Forecasting of sales by using fusion of machine learning techniques”, 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), 2017.