**CSC-433 PROJECT SUBMISSION**

Data Mining

on

Big-Mart Sales Dataset

(Technical Report)

by

**Pradeep Sathyamurthy**

Under the guidance of: Prof. Steve D. Jost

DePaul University

08th June 2017

# Project Summary:

## Introduction:

I have registered myself with Kaggle and AV (Analytic Vidhya) where lots of Machine Learning competitions are held. As part of AV they had opened a challenge to predict sales generated by a retail store called Big-Mart with dataset which contains properties of store and product being sold there. I would like to start with simple linear regression as part of this project scope and later advance with other machine learning algorithm to enhance my output in explaining the variance of sales with respective to the dataset provided.

## 

## Project Scope:

In this project my main interest is to apply the knowledge obtained from courses below to perform data mining on the Big-Mart dataset:

*CSC-433: Scripting for Data Analysis*

*CSC-423: Data Analysis and Regression*

*CSC-465: Data Visualization*

I will restrict my analysis with doing Exploratory Data Analysis (EDA), Treating missing values with decision tree and mean value imputation for factor and numerical variable respectively and finally apply simple OLS linear regression model on the cleaned dataset and observe the variance explained. I am planning to continue this project as part of my Summer courses where I would study about few regularization techniques like Ridge Regression and another advance data mining techniques like Random Forest, XGBOOST, etc., to make my model more sophisticated in explaining high variance with model being free from overfit or underfit.

## Dataset Description:

The data scientists at Big-Mart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a store. Using this model, Big-Mart will try to understand the properties of products and stores which play a key role in increasing sales. This dataset is available on registration to participate in the hackathon conducted by AV through this link <https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/>

***Files:***

We were shared with two files train.csv and test.csv. Former to train the algorithm and later to validate the final algorithm built which do not have the values for dependent variable (count) for competition evaluation purpose by AV. However, for this project since we are requested to show and prove the final model behaviour, we have ***considered only train.csv file*** which has the information of dependent variable. Based on this train.csv file only ***we created our Train [6113 records] and Test data [2410 records]*** for model building and model validation respectively. Thus, for this project purpose test.csv file shared by AV has been discarded. Below are features descriptions available as part of the dataset:

|  |  |  |
| --- | --- | --- |
| SI.NO | Variable Name | Description |
| 1 | **Item\_Identifier** | Unique product ID |
| 2 | **Item\_Weight** | Weight of product |
| 3 | **Item\_Fat\_Content** | Whether the product is low fat or not |
| 4 | **Item\_Visibility** | The % of total display area of all products in a store allocated to the product |
| 5 | **Item\_Type** | The category to which the product belongs |
| 6 | **Item\_MRP** | Maximum Retail Price (list price) of the product |
| 7 | **Outlet\_Identifier** | Unique store ID |
| 8 | **Outlet\_Establishment\_Year** | The year in which store was established |
| 9 | **Outlet\_Size** | The size of the store in terms of ground area covered |
| 10 | **Outlet\_Location\_Type** | The type of city in which the store is located |
| 11 | **Outlet\_Type** | Whether the outlet is just a grocery store or some sort of supermarket |
| 12 | **Item\_Outlet\_Sales *(Dependent Var)*** | Sales of the product in the store. This is the outcome variable to be predicted. |

# Project Methodology Layout

START-EDA

Clean Up

END

Data Split

Final Prediction Model

Model Validation

STOP

Testing Data

Potential Model Problems and Solution

Model Assumption

Model Adequacy

Training Data

Model Building

# Exploratory Data Analysis

R being built above the ancient language, data being in either Factor or Numeric would serve the purpose better. is comfortable with, thus we explored our dataset initially and tried to convert them into Factors and Numeric.

|  |  |
| --- | --- |
| Before Datatype Conversion | After Datatype conversion and column pruning |
|  |  |

It is always wise to start a model building with hypothesis generation and an exploratory data analysis. This will help us to:

* Understand the relationship between the variables
* Gain domain expertise
* Avoid bias based samples
* Build a structure modelling with a structured approach*.*

Best approach to validate these hypothesis is through visualization, below are few EDA[[R]](#_R-Code_for_Bike) done on ***train.csv*** file:

## EDA on Factor variables:

|  |  |
| --- | --- |
| Exploratory Data Analysis on Factor Variables Before Clean-up and Split | |
|  |  |
|  |  |
|  |  |
|  |  |

## Hypothesis:

1. Low fat food is being purchased more compare to the regular fat foods
2. Food products like Fruits and Vegetables, snacks have higher sale; Households, canned, dairy and baking good have average sales and others are bought even less
3. OUT010 and OUT019 have lowest sale compare to others
4. Big mart owns Small and medium sized outlets more when compare to High size outlet
5. Big mart outlets are situated more in Tier3 and Tier2 locations when compare to Tier1 regions
6. Other than 1997, we could see a constant sale obtained in all years till 2009

## EDA on Factor variables:

|  |  |
| --- | --- |
| Exploratory Data Analysis on Numerical Variables Before Clean-up and Split | |
|  |  |
|  |  |

## Hypothesis:

1. Item weight has a normal distribution, which means product of all weight are available in store at equal proportion, it not just the whole sale which is happening in store
2. Product visibility is skewed to right, stores have more of small display area for product more and interestingly there is a size 0 which can be even online sold product
3. MRP of the product is also quite normally distributed, which means product of all price range from $31 to $266 is available in store in equal proportion, so it targets all kind of customers for its sales
4. Total sale revenue is skewed to right, meaning store constantly generate revenue of range $800 to $3000 in each of its outlet mostly

Thus, we hypothesized few scenarios based on our dataset, I would like to highlight same below:

*Groceries like fruit, vegetables and snacks with low fat content with minimum product visibility in a small and medium sized outlet situated in Tire-3 and Tier-2 region should have a comparatively good sale excluding the outlets OUT010 and OUT019*

# Missing Value Treatment:

I tried to look at all variables with values being Na, NaN, NULL and blank. Through below visualization I was able to figure out there are 2 variables with missing values:

1. Outlet\_Size
2. Item\_Weight

|  |  |
| --- | --- |
| Missing values in Factor Variables | Missing values in Numerical Variables |
|  |  |

## Treating the Factor and Numerical Variable:

Outlet\_Size is a factor variable, hence I decided to use decision tree though in CSC-465 to compute the missing values in the dataset. Thus, using below decision tree we imputed stores with outlet identifier OUT018, OUT027 and OUT049 as Medium type and if OUT013 it being of size High and rest all others as small.

|  |
| --- |
| Decision Tree |
|  |

Similarly, I treated the numerical variable Item\_Weight using mean value of the them in a iterative state there by both median and mean was staying close to each other causing no bias in data. Post this I checked for missing values:

|  |  |
| --- | --- |
| Missing values in Factor Variables | Missing values in Numerical Variables |
|  |  |

# Data Split

In this dataset, we totally have 7 factor variables with 4 variables being a quantitative data. So, totally we have 11 variables (7 factor + 4 quantitative). As per thumb of rule we need to have at least 110 sample records to split the data in to train and test data. Since, we have totally 8523 records as part of our train.csv we can do a split[[R]](#_R-Code_for_Bike) of ***Training-Data: Testing-Data = 80:20 ratios*** through which ***Training data can be used to build our model while Testing Data can be used to test our model for model validation and prediction***.

## Training Data:

From here on we will refer our Training dataset with name ***data.train*** which is 80% of Simple Random Sampled data from train.csv file. Data.train has a total number of 8710 samples through which we will be building and training our model.

*Training Dataset: data.train = 80% of (8523 samples of Mart\_train.csv file)*

## Test Data:

From here on we will refer our Testing dataset with name ***data.test*** which is the remaining 20% of Simple Random Sampled data from train.csv file. Data.test has a total number of 1703 samples through which we will be validating our model and subject it for prediction. Since, our testing dataset is a remaining sample left out by training data, ***our data.test is no way a subset of data.train these are two simple random samples of train.csv***.

*Testing Dataset: data.test = 20% of (remaining 8523 samples of Mart\_train.csv file)*

## Exploratory Data Analysis on data.train after clean-up:

|  |  |
| --- | --- |
| Exploratory Data Analysis on Factor Variables After Clean-up and Split | |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
| Exploratory Data Analysis on Numerical Variables After Clean-up and Split | |
|  |  |
|  |  |

From above graphs we could infer that data clean up on factor variable has been done to derive variables with minimum class labels and clean up on Numerical variables either turned the data normally distributed to certain extent. Now we can take this dataset for Model building.

# Model Building and Adequacy:

|  |
| --- |
|  |
| * Over all model was significant with p-value < 0.01 * No significant correlation exists with all numerical variables available * Variables Item\_Fat\_Content, Outlet\_Identifier and Item\_MRP seem significant variables with respective p-value < 0.05 * However, ***the variance that this model could explain was only 56.57% which is considered to be not efficient*** * We applied even the stepwise algorithm which even explains only 56.66% of variance in sales. |

# Model Assumption

We start to create any models with few assumptions, in which two major assumptions are:

1. **All pairs of error terms are not correlated that is error terms are independent to each other**
2. **Error is normally distributed with mean=0 and Standard Deviation being constant**

So, to re-confirm that our assumptions hold good for the model that would be subjected to prediction, we need to perform few residual analyses before concluding the final model.

|  |
| --- |
| Model: Item\_Outlet\_Sales ~ Outlet\_Identifier + Item\_MRP |
|  |

|  |
| --- |
| Residual Analysis  1. Graph plotted between Residual and Fitted is used to confirm assumption-1 2. We expect residual plot with no trends or pattern. From above figure we could infer that ***there is a concrete trend that exists*** in this plot. 3. ***There is a dramatic increase in variability***  Heteroscedasticity:  1. We say a model as heteroscedastic when there is no constant variance. Funnel shape of residual plot clearly identifies the model is heteroscedastic. 2. Even in our residual graph we see a funnel shape and can say that ***our model is heteroscedastic*** in nature.  Normal Probability Plot:  1. From normality plot for the residual, we can notice that most of the points fall reasonably close to straight line which indicates that ***normality assumption is satisfied***.  Outlier and influential Points:  1. Residual vs Leverage graph infers that there are few influential or outliers present. 2. Hence, we calculated for observations which are considered as outliers based on model built, any studentized residual greater than 3 or less than -3 where considered as outliers. We obtained a list of 74 observations. 3. We also wanted to find observations which are influential based on H-hat method. We got a cut off of 0.01319648 and hence considered any value above this as influential point. However, our model fetched only one observation. 4. We compared list of outliers with influential point and there was just 1 matching record in observation 831 which was removed. While the looking at other observations we can concluded that rest of the 73 observations obtained are ***natural outliers and removing them will either over-fit or under fit the model***. |

# Potential Model Problem and Solution:

Above residual analysis clearly indicates that our model is suffering from heteroscedasticity, that is a state with non-constant variance. Thus, we cannot use this model directly for prediction or model validation. We need to fix this. One possible solution is to try transforming the dependent variable and see how our model behaves with respective to explaining variance and residual behaviour. So, we tried to perform model transformation on model selected from model adequacy for residual analysis[[M]](#_Model_Subjected_to).

## Model Transformation:

Since the normal probability plot show a s-shape, we assumed log transformation will do a great difference. Further based on EDA[[E]](#_Exploratory_Data_Analysis) done, we found that **doing log transformation on count had a significant impact** on outliers/influential points and changed the whole shape of box-plot. This confirmed us to try with log transformation for our model[[M]](#_Model_Subjected_to).

|  |
| --- |
|  |
| * Over all model was significant with p-value < 0.01 * Predictors Outlet\_Identifier and ITEM\_MRP is the only variable being significant with p-value<0.05 * Though there is a ***improvement in Adjusted R-Square value from 0.5667 to 0.7241, due to less variable explaining the Sales, we might get Rank issue and data might underfit when subjected to other dataset.*** * Though we assure efficiency into this model using transformation, it should be deciving which can be concluded with model validation. However, after transformation this model ***explains about 71.41% of variance*** in sales data of Big-Mart store. |

## Residual Analysis on Transformed Model

|  |
| --- |
| Model: log(Item\_Outlet\_Sales) ~ Outlet\_Identifier + Item\_MRP |
|  |

|  |
| --- |
| * From above residual graph we can infer that there I ***no more pattern exist*** which indicates no correlation with Residual and Fitted. This ***satisfies our assumption-1*** of error terms to be independent to each other * From the same residual plot, we also see the ***funnel shape no more exists*** and hence can be proved that ***model is no more suffering from Heteroscedasticity***. Model is now Homoscedastic. * S-shape in normal probability plot is also corrected to some extent which means our error terms are normally distributed. This ***satisfies our assumption-2*** of error being normally distributed. * Check on outlier and Influential point was also done which again proved that they ***are natural outliers*** in the system and can be treated as it is in the data. |

# Model Validation:

This is the final stage in building an analytical model. This validation will confirm the following:

1. Is the model over fitting?
2. Is the model under fitting?
3. Is the model good for all samples of the population?

In order to test our model log transformed model[[M]](#_Model_Transformation:) we will be using the data.test which we had split as part of initial data split[[S]](#_Data_Split).

|  |
| --- |
|  |

## Model Performance through Prediction

As part of this process we will inject our data.test into our log transformed model and look at its prediction. Since, we have done log transformation on dependent variable “count”, ***output from prediction needs to be subjected to exponent to get the final prediction value for count variable***. Below are the graphs of prediction performance of our model without exponentiation and with exponentiation.

|  |  |
| --- | --- |
| Prediction Performance on Log Transformed Model | Prediction Performance on model after exponentiation |
|  |  |
| * Above graph in red shows the model performance for the log transformed model * While graph in blue shows the model performance for final model which was re-tansformed * Final model performance proves that ***though the final model explains 72% of variance it behaves badly when it comes for the prediction of Sales data.*** | |

## Model Prediction with Confidence and Prediction Interval values:

Below is the snapshot of model performance, it consists of below details:

1. Actual Count Value
2. Predicted Count Value
3. Predicted Interval Low
4. Predicted Interval High
5. Confidence Interval Low
6. Confidence Interval High

|  |
| --- |
|  |
| * Box in green color highlights the ***closely matched values*** * Box in red color highlights the count values falling into the closest prediction interval range * However, this model is seriously underfitting the data, which cannot be applied for prediction. |

## Final Proposed Model out of regression but this is subjected to further research and enhancement:

|  |
| --- |
| ***Final Model to obtain a regression equation:*** |
| log(Item\_Outlet\_Sales) ~ Outlet\_Identifier + Item\_MRP |
| * This model is seriously underfitting the data, which cannot be applied and advised for prediction. * ***I would like to subject this dataset with other algorithms like factor analysis, SVM, Random Forest and XGBOOST to observe if I can get an optimized result out of it.*** |

# Timetable

|  |  |  |
| --- | --- | --- |
| **Phases** | **Description of Work** | **Start and End Dates** |
| **Phase One** | Obtaining Dataset from Kaggle | 27-May-2017 to 28-May-2017 |
| **Phase Two** | Performing EDA on Big-Mart Dataset | 27-May-2017 to 31-May-2017 |
| **Phase Three** | Basic Model Building | 01-Jun-2017 to 02-Jun-2017 |
| **Phase Four** | Testing and Validation | 03-Jun-2017 to 04-Jun-2017 |
| **Phase Five** | Report Writing and Review | 05-Jun-2017 to 08-Jun-2017 |
| **Phase Six** | Deliverable Submission | 08-Jun-2017 |
|  |  |  |
|  |  |  |
|  |  |  |

# Key Personnel

|  |  |
| --- | --- |
| Team Member | Pradeep Sathyamurthy |
| Professor | Prof. Steve D. Jost |
| Project for | CSC-433 |
| Target Team | DePaul CDM |

# Deliverables

|  |  |  |
| --- | --- | --- |
| Final Report | **Prady\_CSC\_423\_Technical\_Report.pdf** | Contains final Technical Report |
| Raw Data Set | **train.csv** | Raw Dataset downloaded from Kaggle |
| R | **Prady\_Source\_Files\_Bike\_Share.R** | Source File to Run through |
| R\_Data\_Files | **Prady\_Project\_All\_Outcomes.RData** | Can be loaded in R to test all o/p |
|  |  |  |