

Capstone project data science

BigMart-Store Sales Prediction



March 20, 2019

ItGuts

Delhi

Table of Contents

[Titanic Survival Project 2](#_Toc4175028)

[Steps:- 2](#_Toc4175029)

[1. Business Problem Understanding 2](#_Toc4175030)

[2. Data Collection 2](#_Toc4175031)

[3. First Prediction Model 2](#_Toc4175032)

[4. Data Preparation/Cleaning 2](#_Toc4175033)

[5. Data Exploration 2](#_Toc4175034)

[6. Feature Engineering 2](#_Toc4175035)

[7. Model Building by applying multiple machine learning algorithms 2](#_Toc4175036)

[8. Model Evaluation 2](#_Toc4175037)

[9. Model Tuning 2](#_Toc4175038)

[10. Final Conclusion and deployment 2](#_Toc4175039)

# Titanic Survival Project

## Steps:-

### Business Problem Understanding

### Data Collection

### First Prediction Model

### Data Preparation/Cleaning

### Data Exploration

### Feature Engineering

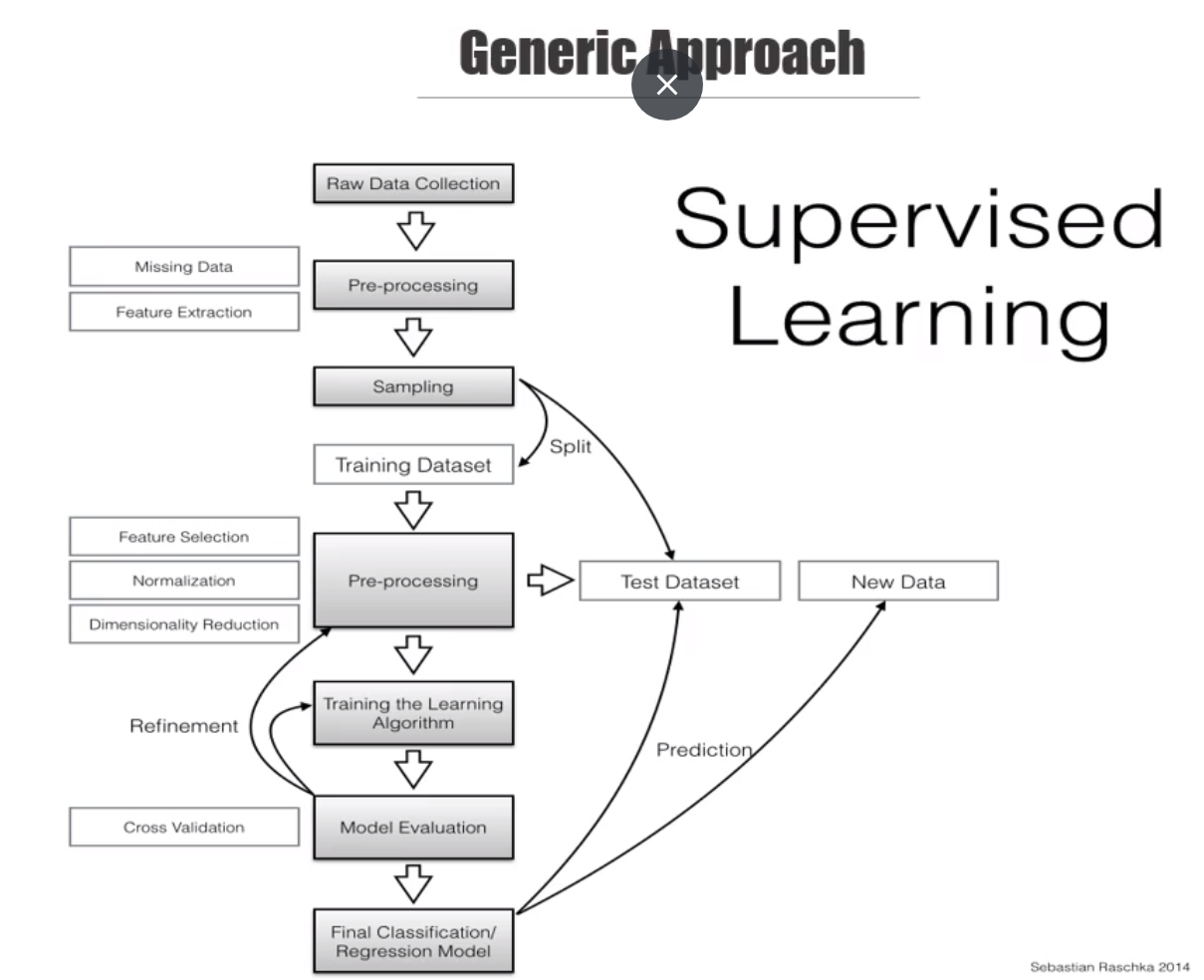
### Model Building by applying multiple machine learning algorithms

### Model Evaluation

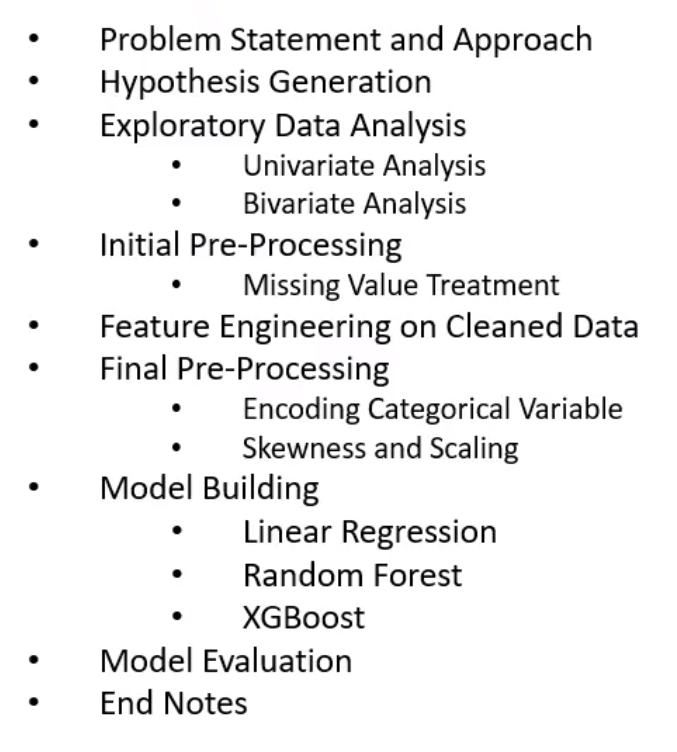
### Model Tuning

### Final Conclusion and deployment

#### Generic Approach



#### Usecase



#### Business Problem Understanding:-

BigMart has collected the sales data of 10 different stores of different cities for 1559 products. In addition, certain attributed of each product and stores have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.

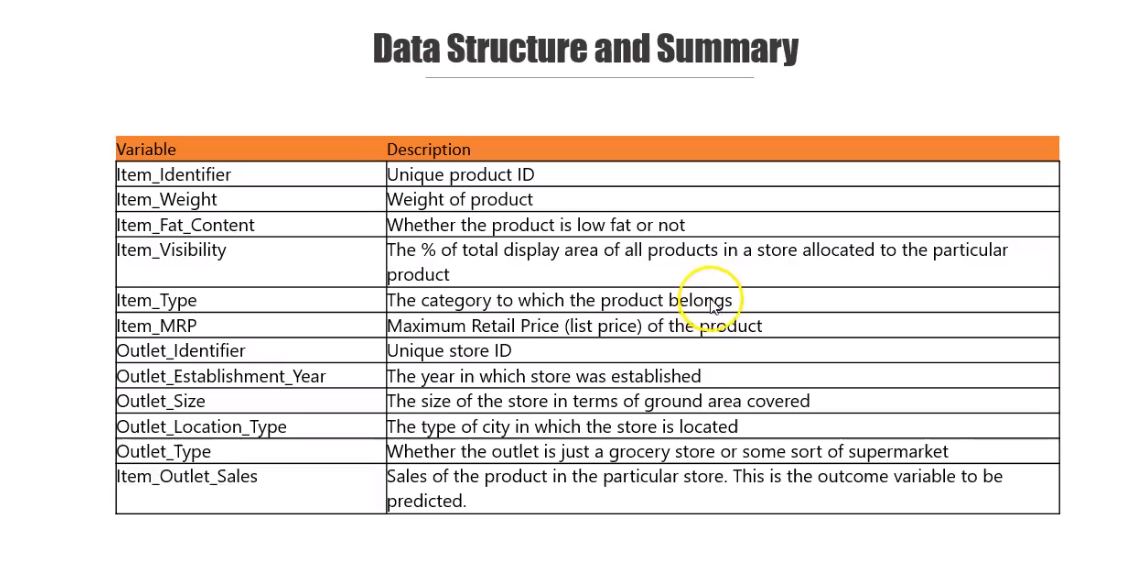
Using this model, BigMart will try to understand the properties of products and stores, which play a key role in increasing sales.

#### Data Collection



#### Data Structure

It has 12 features/attributes in which **Item\_Outlet\_Sales** is my target/dependent variable and rest are possible independent variables.



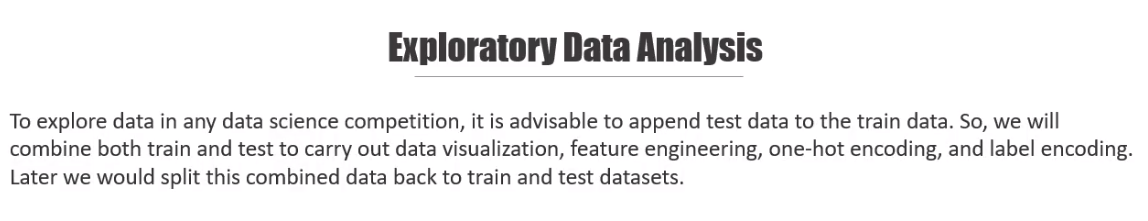
#### First Prediction Model

Without applying any machine learning tools and steps, we can apply our common sense and predict accordingly. It is called as Lazy Prediction.

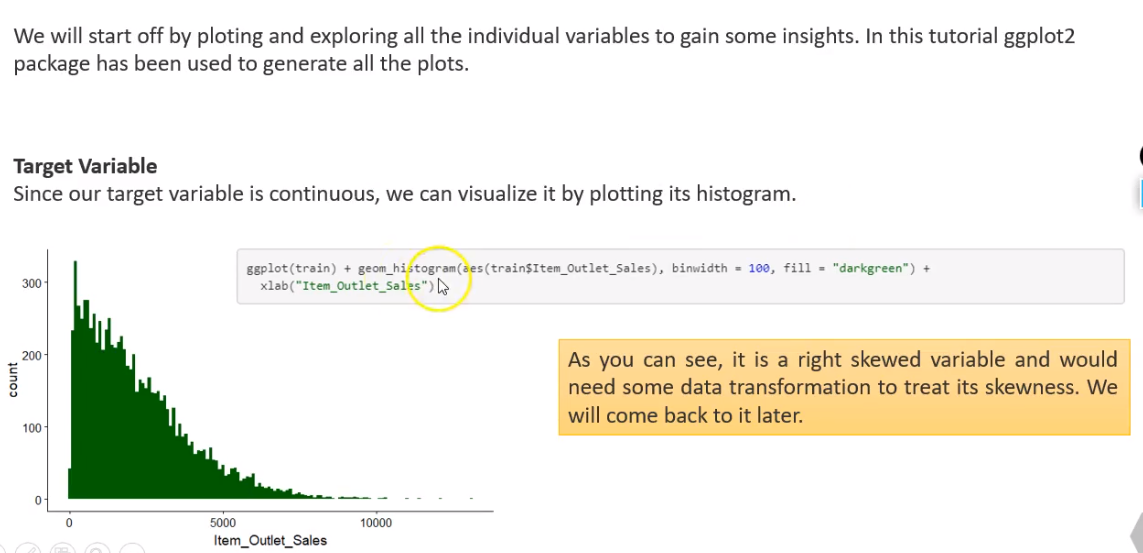
1. Let us assume all the **Item\_Outlet\_Sales** does not depend upon any feature.
2. When we see the sum of **Item\_Outlet\_Sales** based on **Item\_Fat\_Content** I find that the store which have products with less fat content has more sale capacity. So lazily, I predict that to increase sale, store should have low fat content products.
3. I also noticed that the store having products with less MRP has high sales.
4. Here in this assignment, we have nothing to predict, we will see the correlation and variance of the various independent variables that boost the sales.

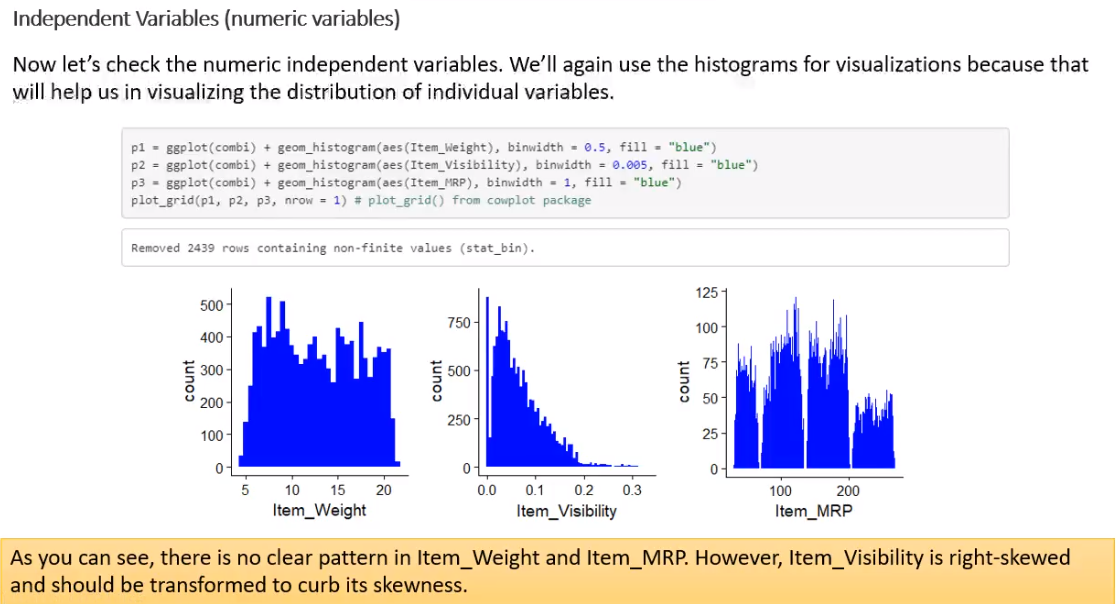
#### Data Exploration

It is also known as EDA. Need to analysis each parameter that how they are contributing the Sales.

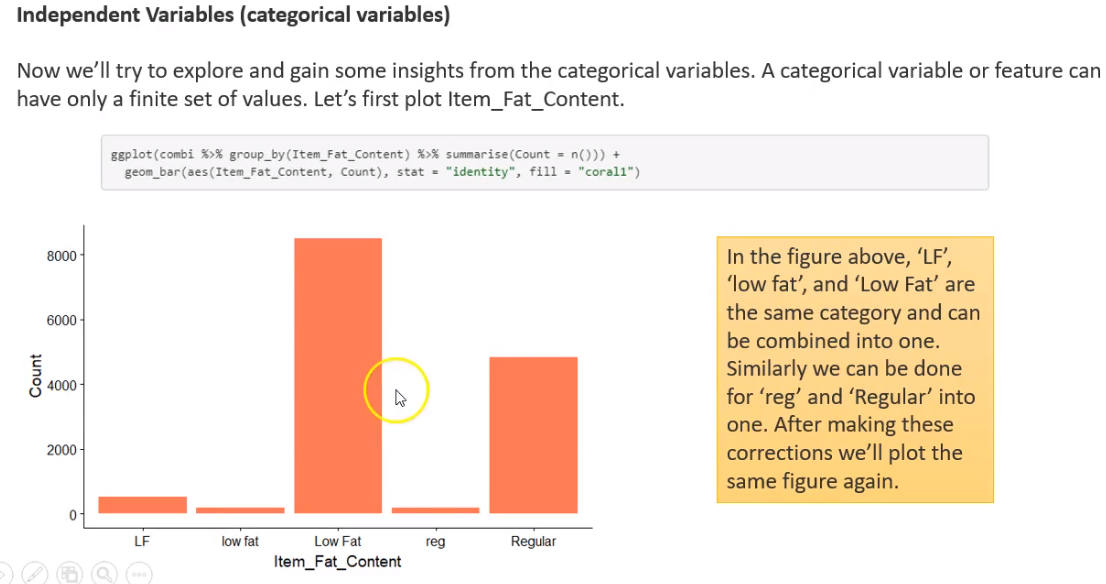


1. Univariate Analysis

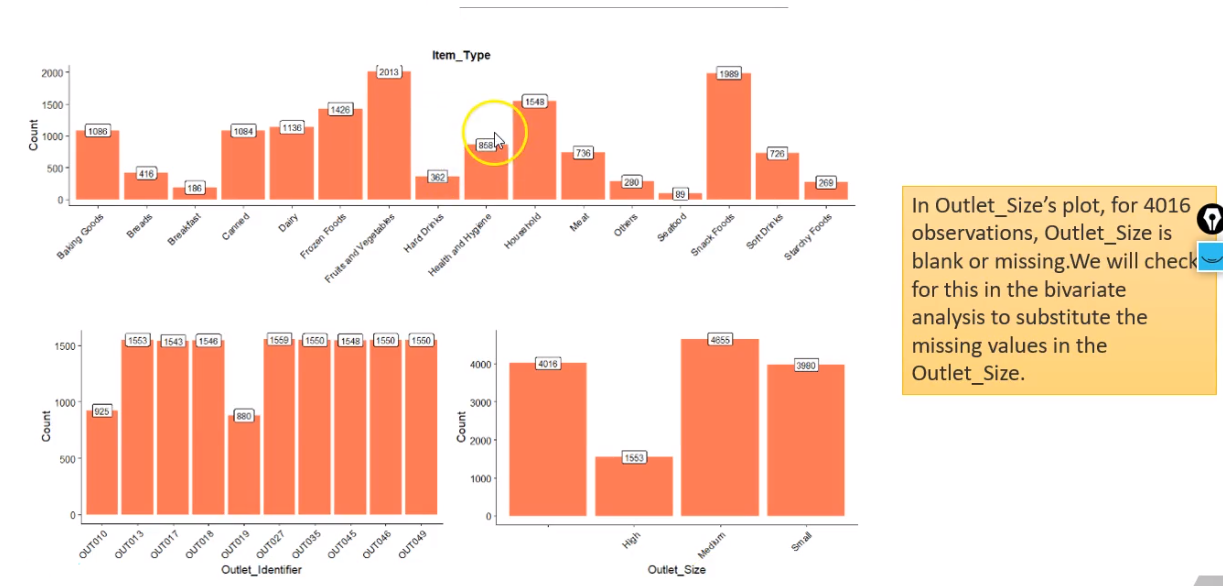




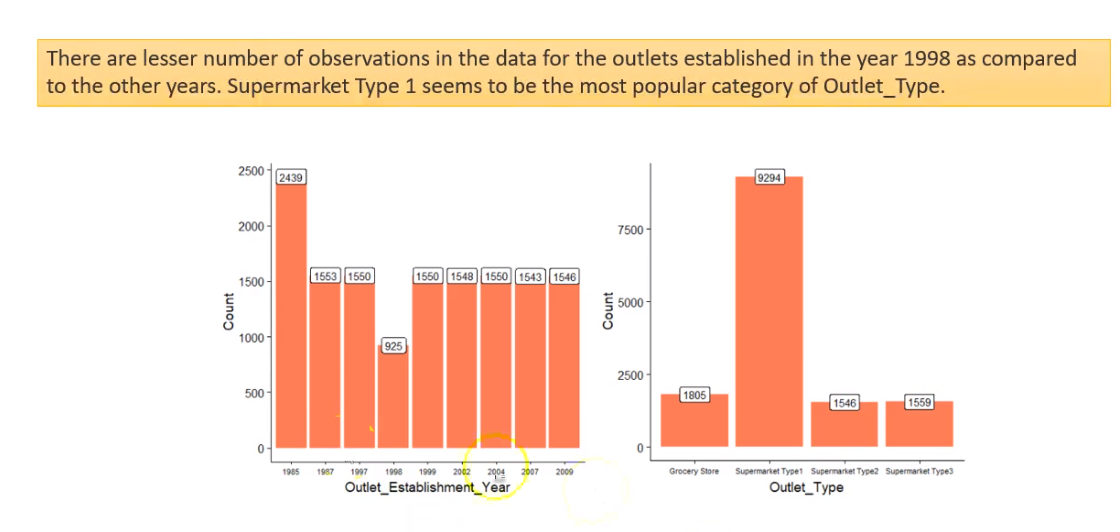
Categorical Variable- Univariate



Low fat products are sold more.



Outliet\_Type and Outlet\_Establishment\_Year



1. Bivariate Analysis

* Item\_Weight vs Item\_Outlet\_Sales
* Item\_Visibility vs Item\_Outlet\_Sales
* Item\_MRP vs Item\_Outlet\_Sales
* Item\_Type vs Item\_Outlet\_Sales
* Item\_Fat\_Content vs Item\_Outlet\_Sales
* Outlet\_Identifier vs Item\_Outlet\_Sales

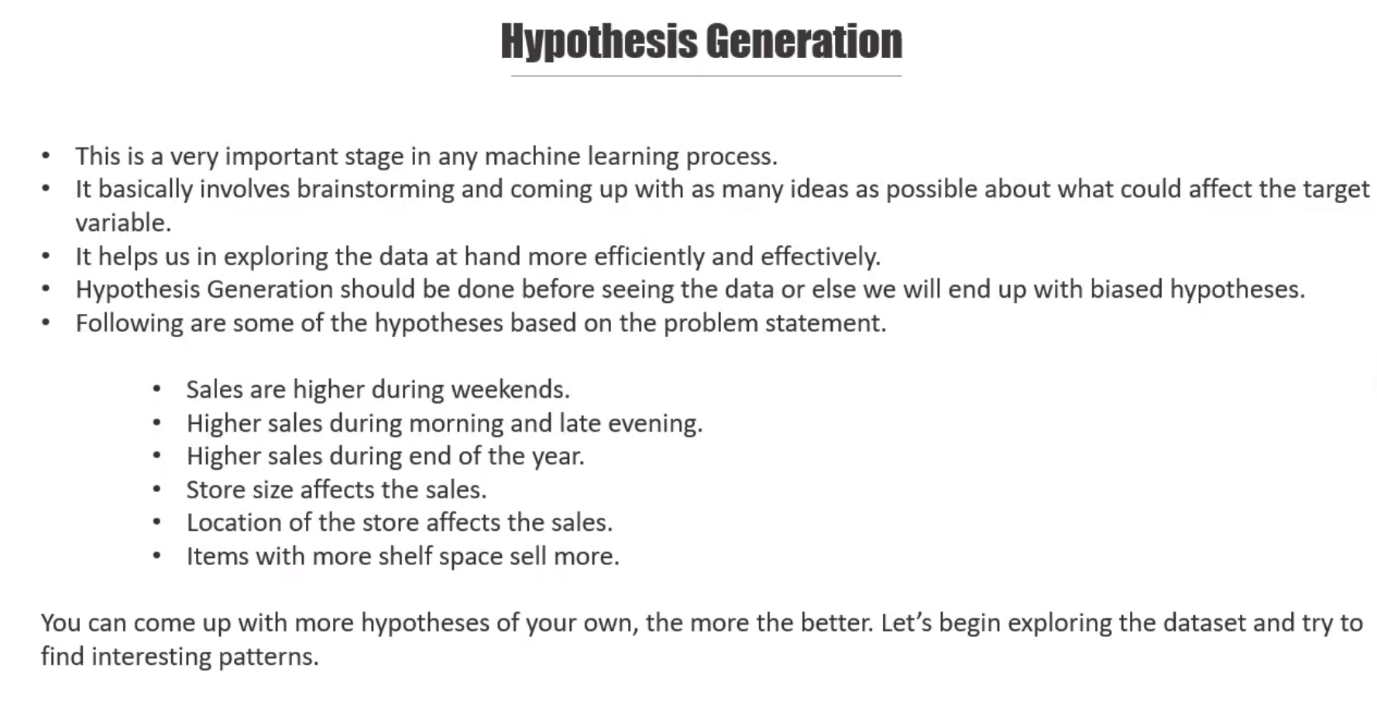
1. Multivariate analysis

Corr plot

Code -



#### Hypothesis Generation

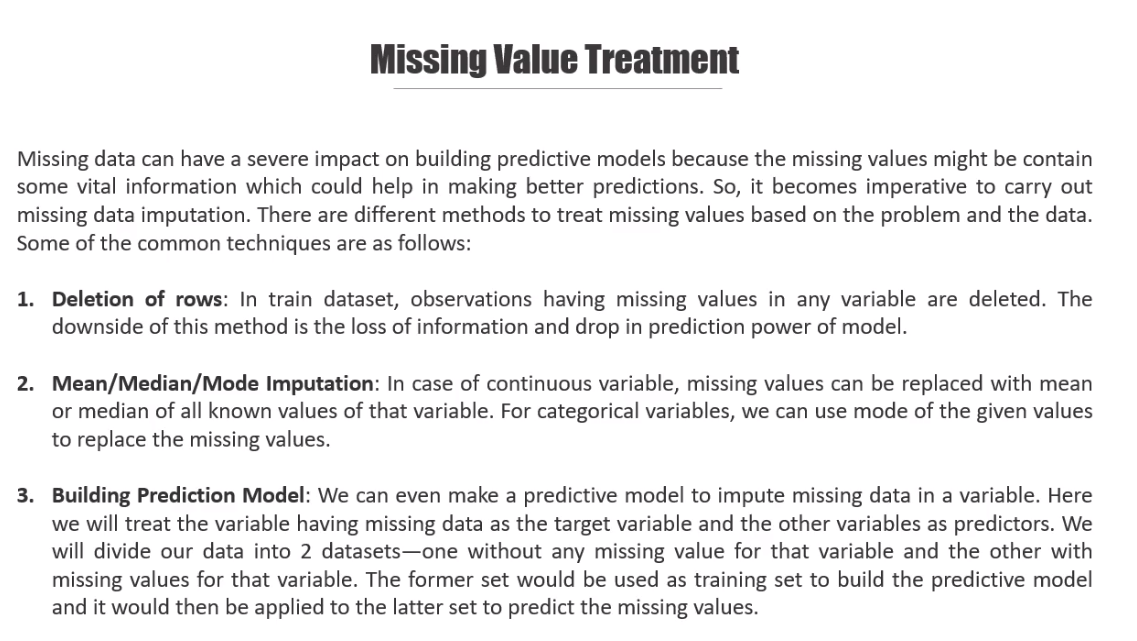


#### Data Preparation

We should consider below factors:-

1. Missing value imputation:-

If missing value of a parameter/filed/column is less than 20%, we should remove the parameter



colSums(is.na(data))

* Item\_Weight has na values, we can replace with mean of Item\_Weight.
* Item\_Visibility has 0 value, so can be replaced by its mean.

1. Outlier detection:-

It is important to know the outliers, as at the time of imputing missing values, we can select either mean or median. If outliers is high then median is the best option.

1. Imbalanced Data Treatment
2. Data Type Conversion

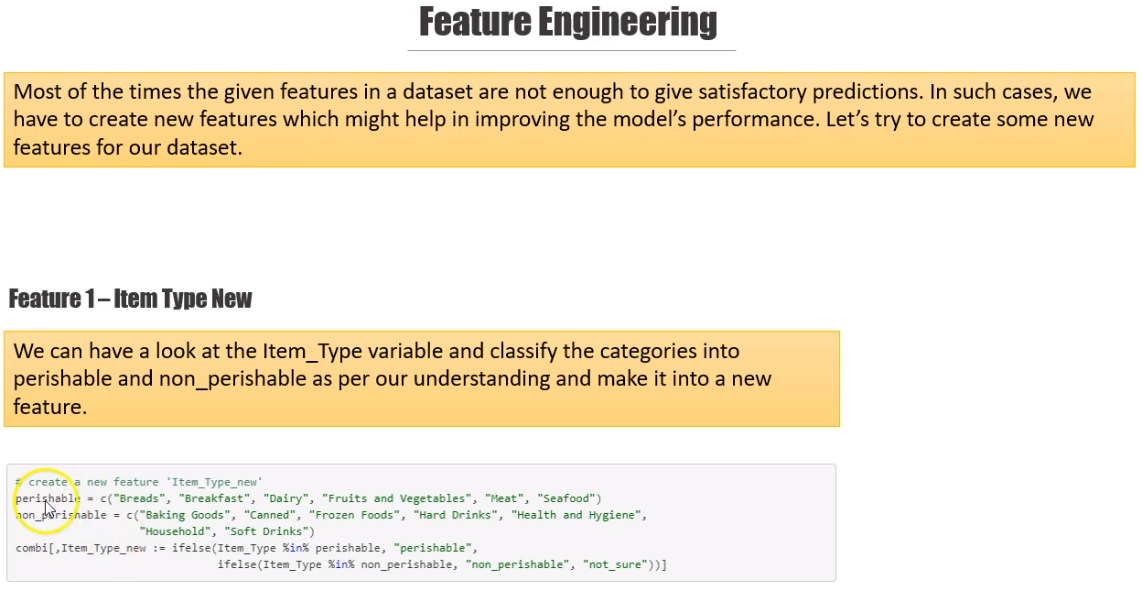
Data Manipulation

Code 

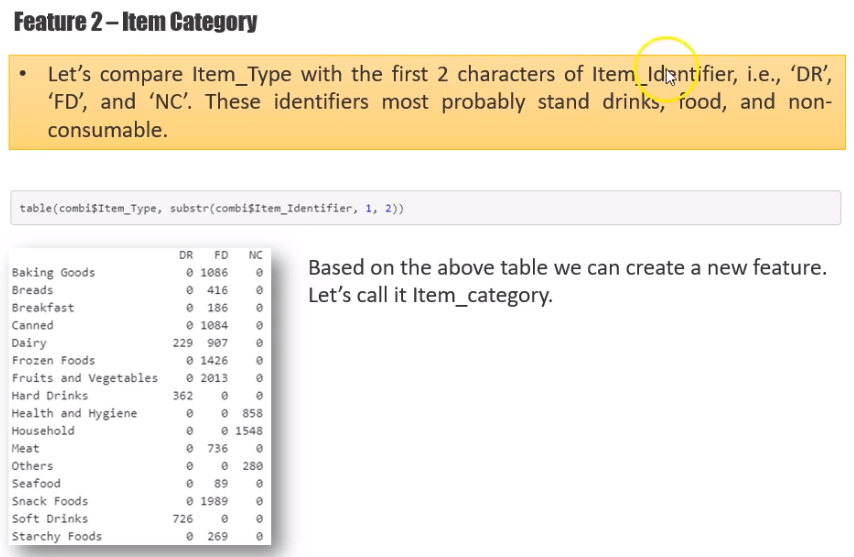
#### Feature Engineering

If you think that by creating new feature/field will help in analysing the fact that would be great for your model.

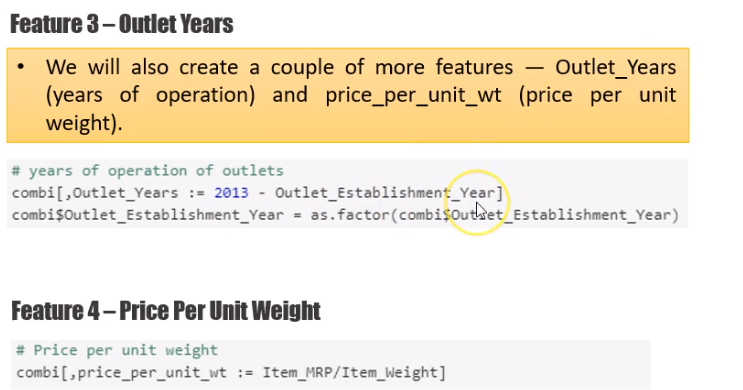
1. Item Type New:- Categorising the product as Perishable and Non-Perishable.



1. Item\_Category:- This can be derived from the first 2 chars of Item\_Identifier. Like FD->Food, NC -> Non-Edible..



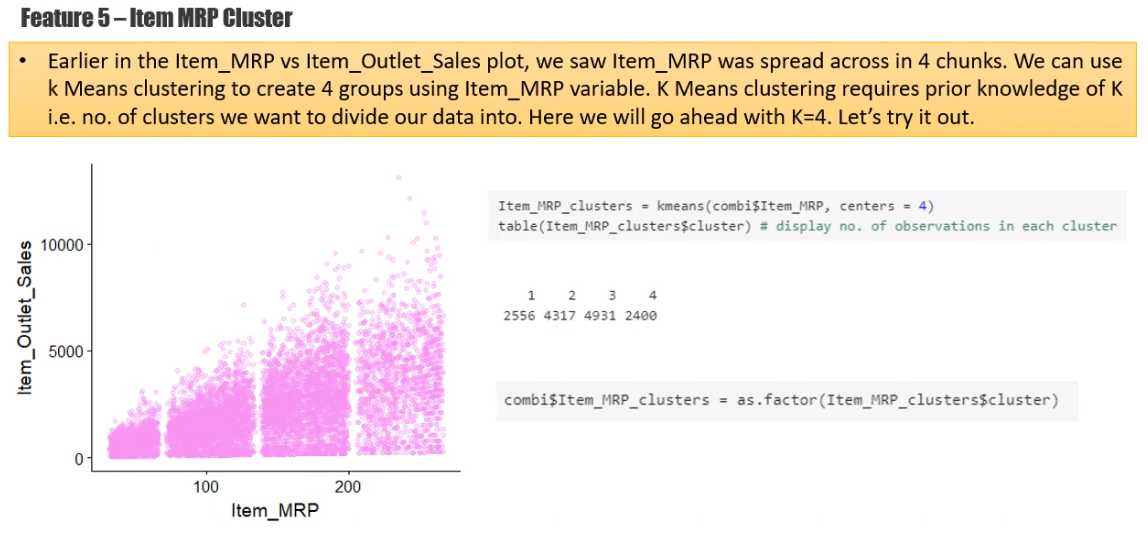
1. Years of operation of the outlet:- 2013 - Outlet\_Establishment\_Year



1. Price per unit weight:-

combi[,price\_per\_unit\_wt := Item\_MRP/Item\_Weight]

1. Item\_MRP\_clusters:- From Item\_MRP, clusters can be created.



Code- 

#### Encoding Categorical Variable- Pre-Processing

Label encoding:-

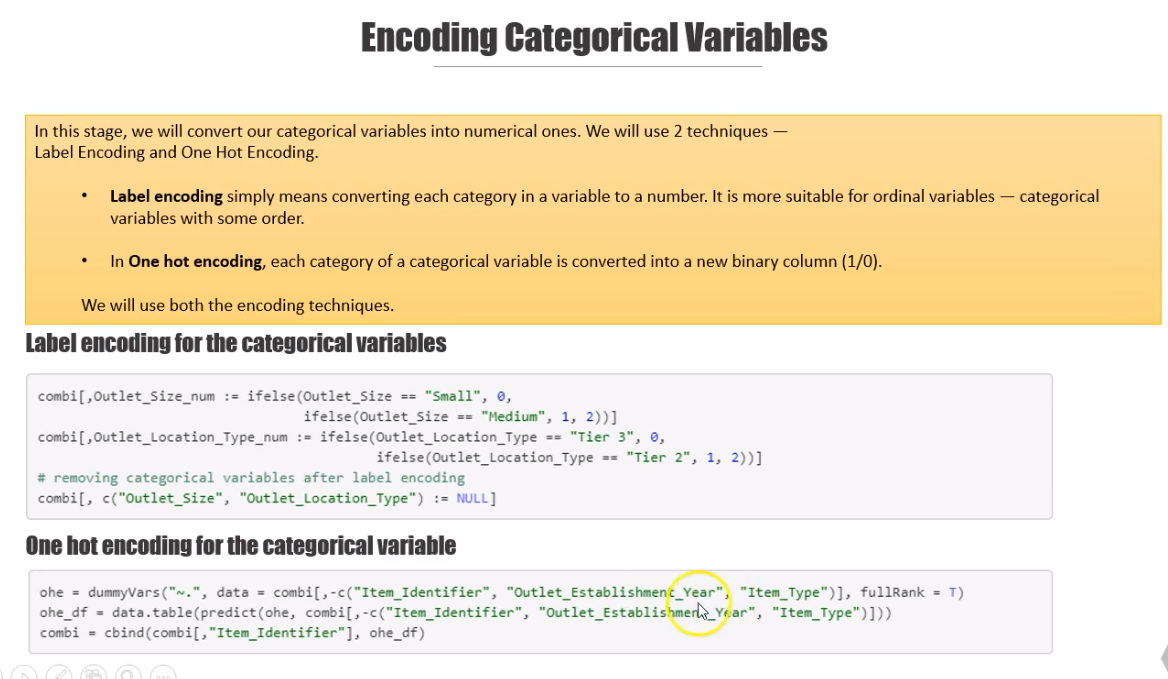
We perform on ordinal variable in which their order has some meaning.

e.g:- Class- Firstclass, secondclass, thridclasss🡪 1,2,3

One hot encoding:-

Here we convert a categorical variable into n-1 new variable, where n is the distinct value.

#e.g Sex has Male, Female,Transgender

# we create two fileds Sex1 and Sex2, if sex1 has 1 then Sex2 will 0(Male), if Sex1 =0 and Sex2=1(Female) and if both are 0 then transgender

Code



#### Skewness and Scaling-Pre-Processing

Scalling is very much important to perform on independent variable if your model uses distance algorithm. Like- SVM, KNN etc..

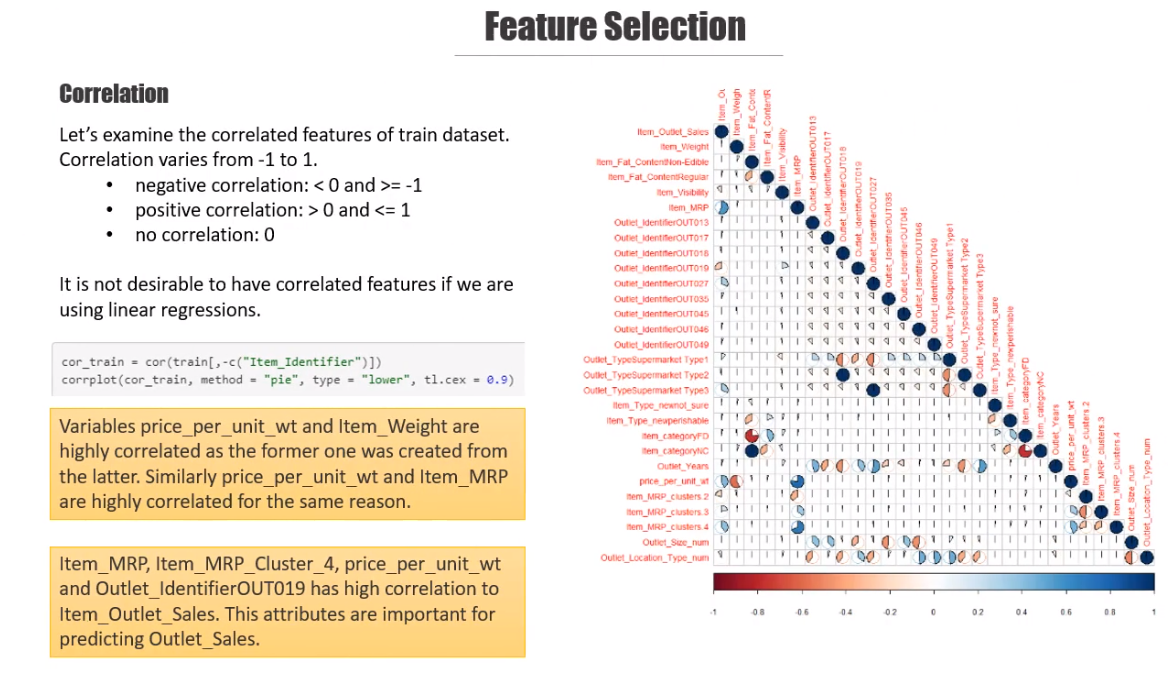


Code :- 

#### Correlation and Feature Selection-Pre-Processing

We check the correlation among independent variables also check multi-collinearity and select the required variables for model preparation.

We leave the target variable and check the collinearity of all independent variables.

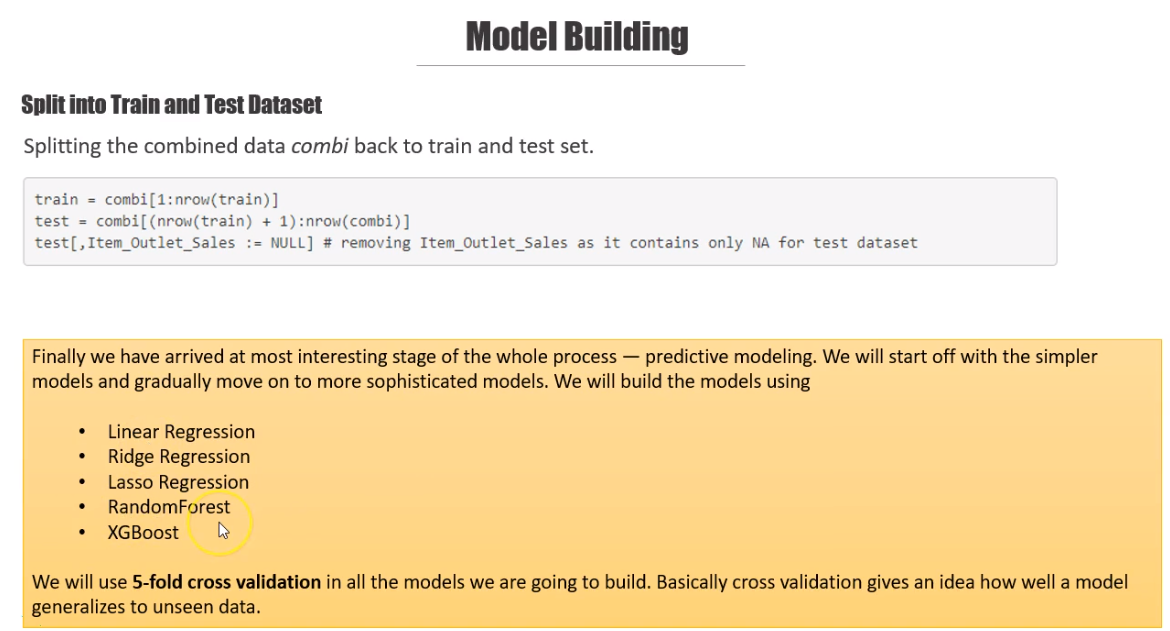


Code 

#### Model Building

We can apply many suitable models after identifying the type of machine learning. This assignment comes under classification supervised machine learning, so we can use logistics, knn, naïve bayes, svm, decision tree or random forest. After that, we can the accuracy and robustness of each model.

We are using Logistic regression, Decision tree and Random forest.



##### Logistic Regression



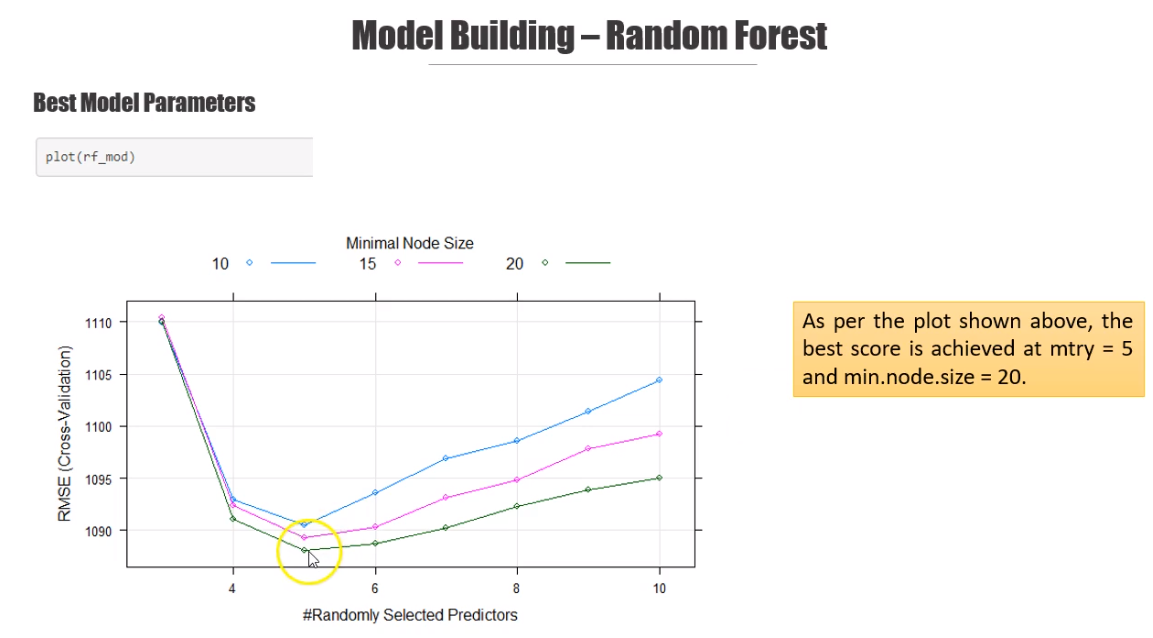
##### Regularised Regression Model



##### Lasso and Ridge



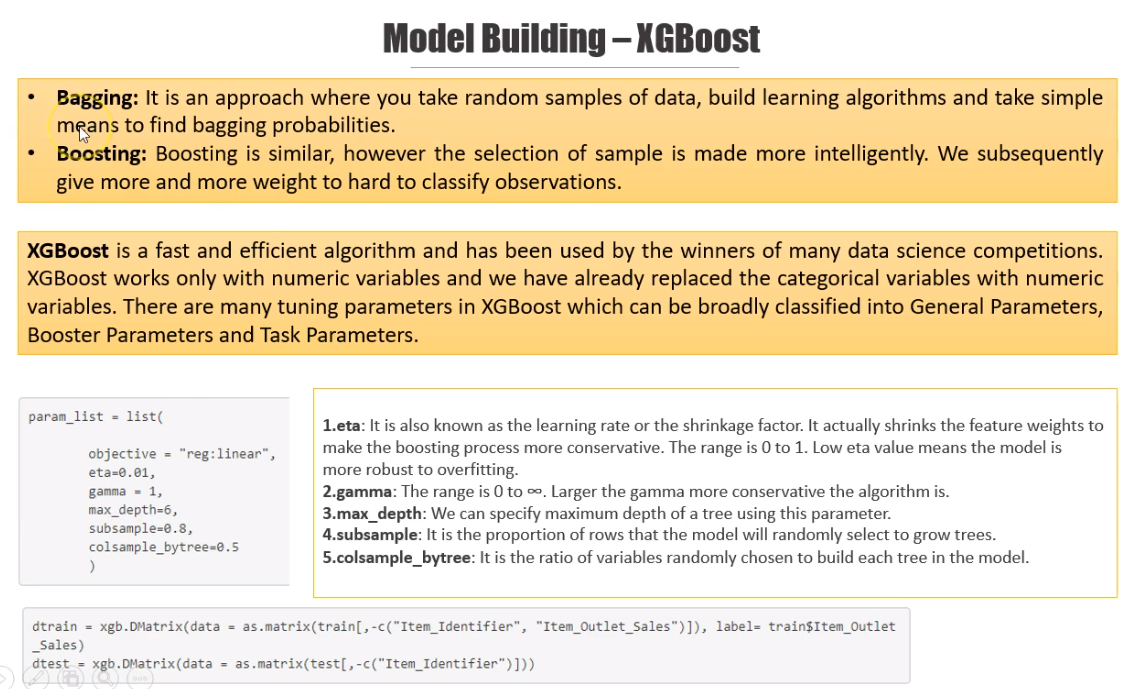
##### Random Forest





##### XGBoost

It is used only for numerical parameters. We have already converted categorical to numerical





Combined:-

