**PROBLEM STATEMENT**

The data scientists at BigMart 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 particular store.

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

**PROBLEM DEFINITION AND OBJECTIVES**

Sales is the key for any business. Supply and demand depends on sales. BigMart have a sales of 1559 products across 10 stores in different cities. Here the aim is to build a predictive model and find out the sales of each product at a particular store.

Before entering into analysis always know about the problem statement to have a basic knowledge about the problem and solution to be achieved.

The purpose of this project is to predict sales in BigMart. The objective of this project is to build a linear model that could successfully determine what would be the sales.

**DATA ANALYSIS**

The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis. Linear model assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). Implementation includes importing necessary libraries, cleaning and analysing the dataset, building various models and using the best model for prediction.

Data analysis contains various steps to be taken care off: Defining the business objective, Sourcing and collecting data, Processing and cleaning the data, Performing exploratory data analysis (EDA), Selecting building and testing the models, Deploying the models, Monitoring and validating against stated objectives.

**EDA CONCLUDING REMARK**

Exploratory Data Analysis is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses. Getting maximum insights from a data set, Uncover underlying structure, Extracting important variables from the dataset, Testing underlying assumptions and determining the optimal factor settings.

After importing libraries and exploring the data, EDA is done for the dataset. It includes analysing the dataset and data explanation. The following are the steps of EDA done to the BigMart dataset:

1. Distribution of target variable, categorical and continuous variables are determined using visualisation techniques. Relationship between features are determined.

2. Remove if there is any duplicate rows.

3. Check for missing values and null values. Treat it using central limit theorem or remove it if filling gives biased result.

4. Check for outliers and skewness. Treat it using z-score or Inter-quartile range.

5. Check for correlation among the features. If presence of correlated variables observed, reconfirm using graph like scatter plot and remove it.

6. Check for linear association between the features(multicollinearity) using variance inflation factor. It is absorbed when two or more features are highly linearly related. Remove if it is present in high amount.

**PRE-PROCESSING/MANIPULATION PIPELINE**

Data preprocessing refer to the manipulation of data before using in machine learning models to ensure high performance. Data preprocessing ensures accuracy, completeness and consistency. Missing data and noisy data are treated. Data cleaning, integration, reduction, transformation are done. Feature selection/ extraction/ reduction is carried out further. The following are the pre-processing steps done to the dataset:

1. Feature reduction is already done through multicollinearity and correlation.

2. Where there are more features which does not give information to predict the target, remove it using Principal Component Analysis to get better performance.

3. Transform the data using any encoding techniques if there is any object type of data.

4. Normalization and standardization of data is necessary to keep all values in same range to get enhanced result.

5. Divide the dataset into target and label. Use train test split to split the data into training and testing data

**BUILDING ML MODELS**

Machine learning involves the use of machine learning algorithms and models. Machine learning (ML) is the study of computer [algorithms](https://en.wikipedia.org/wiki/Algorithm) which are comprised of model data and a prediction algorithm. It is seen as a part of [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence). Machine learning algorithms build a model based on sample data, known as [training data](https://en.wikipedia.org/wiki/Training_data), in order to make predictions or decisions without being explicitly programmed to do so. There are various ML models to predict the target variable using on various techniques.

BigMart dataset is a linear regression problem. Linear Regression has a wide range of real-life applications. It is essential for any machine learning problem that involves continuous numbers – which includes: Financial forecasting (like house price estimates, or stock prices), Sales and promotions forecasting, Testing automobiles, Weather analysis and prediction,  Time series forecasting, etc., The following are the factors to consider the best model:

1. If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression and we use the same. A linear line showing the relationship between the dependent and independent variables is called a regression line. Our main aim is to find the best fit line.

2. Mean-Squared error is the cost function which is the average squared error occurred between the predicted values and actual values.

3. The distance between the predicted value and actual value is called the residual. Here Gradient descent is used to minimize the Mean Squared error.

4. The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called optimization. R-squared is the statistical method that determines the goodness of fit.

**Various ML models to determine the value:**

**Linear Regression:** Supports only linear solutions. Easy and simple implementation, fast training is done. The algorithm assumes the input residuals to be normally distributed and assumes input features to be mutually independent. It is a parametric model.

**Decision Tree:** Decision trees supports non linearity. When there are large number of features with less data-sets(with low noise), linear regressions may outperform Decision trees. In general cases, Decision trees will be having better average accuracy. Decision trees handles colinearity better.

**SVM:** It supports both linear and non-linear solutions using kernel trick. Handles outliers better. It performs well when the training data is less, and there are large number of features.

**KNN:** KNN is a non -parametric model. KNN is slow in real time as it have to keep track of all training data and find the neighbor nodes.

**Random Forest Regression:** Random Forest Regression is a supervised learning algorithm that uses ensemble learning method (ensemble of decision trees) for regression. It is quite a robust algorithm. It has proven to be reliable and effective. Random Forests very often outperform linear regression.

**Cross - Validation**

After using multiple models to determine the target cross-validation is done to verify whether the model is underfitting or overfitting the data. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

**Hyper Parameter Tuning**

Hyper parameter tuning is done with minimum of 4 parameters to get a better tuned result. Hyper parameters are important because they directly control the behavior of the training algorithm and have a significant impact on the performance of the model.

**CONCLUDING REMARKS**

Dataset with 8523 rows and 12 columns is shortlisted to have 6 features and 1 target with standardized data of 8193 rows and loss of data is approximately 4%.

Imported necessary libraries and explored the data. Encoded the data using label encoder and found a column with null values which is not going to support the target with information, so removed it. Checked for correlation and multicollinearity which appeared not so strong. Visualised the continuous variables using distplot and categorical variables using countplot. Checked for outliers and skewness. Removed outliers using IQR since it was in meager amount. The relationship between dependent and independent variables are looking fairly linear. On selecting the best model, tune it using hyper parameter tuning. Post tuning the R-squared value is 92%.

On using this model, determining the sales for future would give 92% true answers on the forecast.