REGRESSION – Made easier…

Synopsis

Note: I made this work as a part of my intern activity. I used various portals to understand the Regression concepts and prepare it.

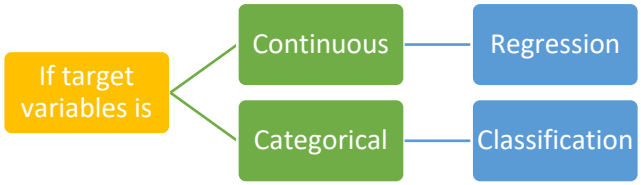
CHAPTER 1:

Introduction to Machine Learning:

Machine learning is a technique which helps the computer to learn and improve from experiences and perform certain activities. It's a part of artificial intelligence.

Machine learning is of three types;

Supervised learning: Data has a target variable- which acts as a mentor for the algorithm during the training phase of the model. It is further classified in to two types based on the task it does.



Unsupervised learning: Data doesn't have target variable. Here the Model finds meaningful insights only from features.

Reinforcement learning: It is based on behavioral psychology.

CHAPTER2:

REGRESSION:

What is Regression?

Regression is a predictive modelling technique. It estimates the relationship between target variable and predictor variables to predict the target variable. When we have to predict the target variable, we go with Regression. Regression is the task of predicting the continuous quantity. A regression algorithm also predicts the discrete value in the form of quantity. Regression is the first technique to be known by any Data Science aspirant.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between target (dependent) variable and predictor (independent) variable to predict the target variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

TYPES OF REGRESSION:

Linear Regression: When the target variable is continuous, we go with linear regression.

Logistic Regression: When the target variable is categorical, we do logistic regression.

TEMPLATE:

Algorithm (explanation-What, Why, Where?)

Various types

Methodology of working (How?)

Assumptions of Algorithm

Issues of Assumptions and Remedial Measures

Measures involved in the algorithm and interpretation

Evaluation metrics

Limitations

Implementation in R / PYTHON / SAS

CHAPTER 3

Linear regression

Types of Linear regression

Methodology: OLS and Gradient Descent

Assumptions of Linear regression

Issues of assumptions & remedial measures

Variable reduction techniques

Measures involved in linear regression & interpretation

Evaluation techniques

Limitations

CHAPTER 3:

LINEAR REGRESSION:

When the target variable is continuous, we go with linear regression. The predictor variables can be continuous or categorical. Linear Regression establishes a relationship between **target variable (Y)** and one or more **predictor variables (X)** using a **best fit straight line** (also known as regression line) to predict the target variable.

There are two types of linear regression based on the number of predictors available.

Simple Linear Regression:

When the dataset has only one predictor variable and a target variable, simple linear regression needs to be done.

E.g.: Imagine we have two variables in our dataset such as speed and distance. Here speed is the predictor, distance is the target variable. It's a simple linear regression problem.

Equation: **Y= mX + c + e**

Y→ target variable

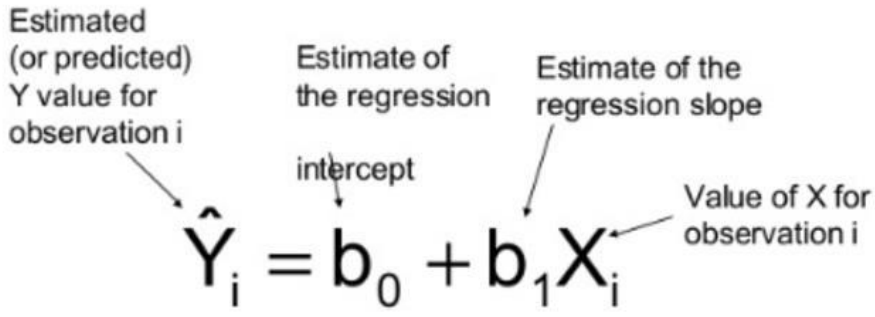
X→ predictor

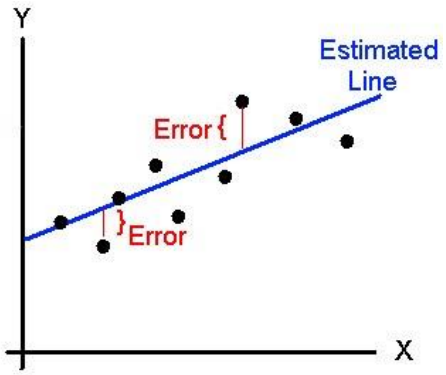
m→ slope→ effect of one variable on the target variable.

c→ intercept → mean value of the variables.

e→ error term or residual

This is a linear equation. This is used to predict the target Y based on the predictor variable X.





Multiple Linear Regression:

When the dataset has two or more predictor variables and a target variable, we do multiple linear regression.

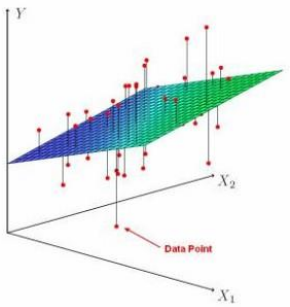
Eg.: Imagine we have dataset which has variables such as Locality, House type, Built-up area, City, Year of construction & House Price. Our task is to predict the house prices in future using the features given. In this dataset house price is the target variable and all other variables are predictors.

Equation: **Y= m1X1 +m2X2 + m3X3 +……+ mnXn + c + e**

Here we have X1, X2…., Xn because the dataset has lot of predictors**.**

Generally multiple linear regression is prone to multicollinearity issues.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n + \varepsilon$$



Techniques to do prediction in linear regression:

Here we find the best fit line in the scattered plot.

Best fit line? It is the line in the scatter plot which better approximates the predictors and the target variable.

How to draw best fit line? There are two techniques available to draw the best fit line

- **Ordinary Least Square Method:** The algorithm will intuitively draw all possible straight lines in the plot. Then finds the difference between actual and predicted values (which is termed as residual). The residuals are squared and added for each possible line separately. At last the line which gives the less value for the above calculation is the best fit line. All these are done by the model behind the scenes. We will get only the best fit line with coefficients.

- **Gradient Descent:** It is one of the best optimization methods that we can use to solve the various machine learning problem. *G*radient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost). In OLS method we try to find the minimum sum of squared error to fit the best fit line. Here, in GD we make hypothesis which wants to give least cost function. The cost function

measures how far my hypothesis is from the actual data. Here the hypothesis means all the possible lines which are drawn iteratively. Lots of iteration are done. At last, the hypothesis which gives the minimum cost function is the optimal line or best fit line.

After fitting the best fit line, prediction is done through **extrapolation and interpolation in that best fit line**. Thus, the target variable Y is predicted for given X. The best fit line forms a linear equation.

OLS vs Gradient Descent:

Ordinary Least Square Method	Gradient Descent
When predictors and observations are less, we use OLS.	When we have many predictors and observations, we use GD.
When the data is linear, we use OLS.	When the data is non-linear and complex, we use GD.
OLS is the best in giving inferences about variables.	When it comes to prediction of high dimension problems, GD is the best.

Assumptions:

There are certain assumptions of linear regression, which the data has to satisfy to apply linear regression on it.

1. There must be **linear relationship** between independent and dependent variables. E.g.: If X increase Y also should increase and vice versa. This is a linear relationship.

2. There should not be multicollinearity, autocorrelation, heteroskedasticity issues.

3. Linear Regression is very sensitive to outliers. Because the regression line gets affected due to outliers present in the dataset. It's advisable to treat outliers in the data preparation stage.

In real life scenarios all the conditions may not be met. So, check whether the assumptions are satisfied. If not assess how to rectify it.

Issues and treatments:

Check for Linearity in data:

In simple linear regression: Plot the target and predictor in the scatter plot. Then find whether it's linear or not.

In multiple linear regression: The residuals should be plotted against each predictor variables. If the data is linearly related, we should not see any pattern in the plot. The reason that we plot residual (instead of target variable) against each independent variable is because, in multiple regression we look at many predictors influencing target variable. If we plot only one predictor and target, we are not looking at the impact of other predictors on target.

Remedial measure for non-linearity:

In case of non-linear relationship, we transform the data in to linear. There are many ways to transform variables to achieve linearity for model such as; taking log, square roots etc.., **Variable Transformations:** Taking log, sqrt, cube root, natural log to transform the variables.

Multicollinearity arises when we have more predictors in the dataset. It refers that there is a chance that one or more predictors might be highly correlated to each other in a dataset. This affects our linear model since highly correlated variables explain same matter in different formats. Thus, model gets confused.

Causes:

It makes some variables statistically insignificant, increase the standard error of coefficients. So, multicollinearity is an issue in regression.

Eg.: Imagine two singers are singing the same song at a time. Then we will feel confused to judge them both since they sing the same song simultaneously in different voices.

How to test Multicollinearity?

We can do pairwise correlation (or)

Variable Inflation factor (VIF) is used to find whether multicollinearity is present or not.
If VIF = 1, no predictors are correlated to each other.
If VIF=1.5 to below 5, it shows predictors are slightly correlated.

If VIF >= 5, Predictors are highly correlated.

If VIF>10, Predictors are very highly correlated

Here the regression coefficients are poorly estimated due to multicollinearity.

DANGER SITUATION

Remedial Measure for Multicollinearity:

Remove highly correlated predictors from the model. If you have two or more factors with a high VIF, remove one from the model. Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared.

Also, various variable reduction techniques can be used to eliminate the problematic predictors.

Autocorrelation occurs when the residuals are not independent from each other. Eg: this typically occurs in stock prices, where the price is not independent from the previous price. It means the error terms (residuals) following a pattern. It's a bad situation for regression model. We can't have a regression if the error terms are correlated and follow a pattern.

Mostly when the predictor variables are time series in nature, we will come across the autocorrelation issue. i.e: error term is correlated to its previous error terms.

This is applicable especially for time series data. Autocorrelation is the correlation of a time Series with lags of itself. When the residuals are autocorrelated, it means that the current value is dependent of the previous (historic) values and that there is a definite unexplained pattern in the Y variable that shows up in the disturbances.

Causes:

In the presence of autocorrelation, the OLS estimators may not be efficient (that is, they may not achieve the smallest variance). In addition, the estimated standard errors of the coefficients are biased, which results in unreliable hypothesis tests (*t*-statistics). The OLS estimates, however, remain unbiased.

How to test Autocorrelation?

Do scatterplot of residuals. If the plot shows a pattern, then autocorrelation is present in the model.

Also, we can use acf() function in R to find whether auto correlation is present or not.

Durbin Watson test is the statistical technique to find autocorrelation. It works well in finding the first order autocorrelation. First order autocorrelation denotes the error term is correlated with its near preceding (just one lag before the error term) error term.

We also have Ljung-Box Q test to test the autocorrelation. Based on the p values of either of these tests we conclude whether autocorrelation exist or not.

E.g:

Null Hypothesis: No autocorrelation exists in residuals.

Alternative Hypothesis: Autocorrelation exists in residuals.

If p value <= alpha, we reject null hypothesis

If p value> alpha, we accept null hypothesis.

Remedial Measure for autocorrelation:

Omit the predictors one by one iteratively and check for autocorrelation.

Some variable transformation techniques can be done.

Heteroskedasticity, (the violation of homoscedasticity). The problem that heteroskedasticity presents for regression models is simple. Recall that ordinary least-squares (OLS) regression seeks to minimize residuals and in turn produce the smallest possible standard errors. By definition, OLS regression gives equal weight to all observations, but when heteroskedasticity is present, the cases with larger disturbances have more "pull" than other observations.

Homoskedasticity is that the residuals should have constant variance. It occurs more often in datasets that have a large range between the largest and smallest observed values

Let's take a look at a classic example of heteroscedasticity. If you model household consumption based on income, you'll find that the variability in consumption increases as income increases. Lower income households are less variable in absolute terms because they need to focus on necessities and there is less room for different spending habits. Higher income households can purchase a wide variety of luxury items, or not, which results in a broader spread of spending habits.

One more instance is that if my linear model predicts the accidents in various cities. Cities with high population will have high accidents when compared to the cities with low population. This is an example of heteroskedasticity.

This becomes a problem that linear regression assumes that the spread of the residuals is constant across the plot. Heteroskedasticity violates this and model results may become incorrect.

Causes:

Heteroskedasticity makes the model less precise.

Heteroscedasticity tends to produce p-values that are smaller than they should be. This effect occurs because heteroscedasticity increases the variance of the coefficient estimates but the OLS procedure does not detect this increase.

This problem can lead you to conclude that a model term is statistically significant when it is actually not significant. We will have insignificant predictors in our model which is bad sign.

How to find the presence of heteroskedasticity?

To check it, plot residuals against the predicted Y. If the graph shows no pattern then its homoskedasticity. If the graph shows a pattern then heteroskedasticity is present

Remedial Measure for heteroskedasticity:

When there is an evidence of heteroskedasticity; Use of weighted least squares regression would be more appropriate, as it down-weights those observations with larger disturbances.

Weighted regression is a method that assigns each data point a weight based on the variance of its fitted value. Weighted regression minimizes the sum of the weighted squared residuals. When you use the correct weights, heteroscedasticity is replaced by homoscedasticity.

The idea is to give larger weights the points which are close to regression line and small weights to points which are far away from the regression line. (The idea is to give small weights to observations associated with higher variances to shrink their squared residuals). Thus, the variance among the points gets balanced and model becomes robust.

Also, we can transform the variables such as; taking log for the target variable to get rid of heteroskedasticity.

Variable Reduction or Selection Technique in Multiple Regression:

- Domain expertise will help us to reduce variables intuitively. But it will take more time.
- Highly correlated predictors are found and omitted using VIF technique.
- We can use Stepwise Regression. It is a combination of forward and backward selection. It starts with all variables to build a model. Then drop the least significant variable. It is an automatic iterative process used to remove predictor variables. It is fine-tuning the model to choose the best predictor variables from the available options. It helps us to iteratively explore which predictors seem to provide the best fit.
- **Use Partial Least Squares Regression (PLS) or Principal Components Analysis**, to reduce the variables.
- Also, AIC method helps to reduce the predictors. When we have a greater number of variables. We will use AIC function to reduce the number of variables. Because, if we do it manually, it will take more time. These are the two types of linear regression.

Measures involved in Linear Regression and its interpretation:

Residual: The difference between actual and predicted values.

Sum of Squares of Residual (SSR) : Each residual is squared and added. We square the residuals to get rid of sign (+ve or -ve) issues. Then added to get SSR. The lines which gives the least SSR is considered as the best fit line under Ordinary Least Square Method.

The result table has slope values and the intercept, which forms a linear equation and helps to predict the target variable.

Evaluation metrics:

Measures of Model precision:

- **p value** is the probability value found by assuming null hypothesis is true. Based of p value we accept or reject the null hypothesis. Generally, p value shows how significant the variable in linear regression. Predictors which has p value with star ratings (*** ,**, *) in the model result are considered as the highly significant variables. Other predictors are omitted.

- **Coefficient of Determination:**
  R^2 in the model result is an important measure. It gives information about the goodness of fit of a model. At least R^2 value should be 75% for the model to become the best model. It is proportion of variability in Y explained by the regression model.
  R^2 =Square of Correlation Coefficient. I.e. $r^2 = R^2$
  If R^2 >=75% , then there is a strong linear association .

- **Coefficient of Multiple Determination:**
  Adjusted R^2 is used as the measure of goodness of fit when we have multiple predictors. In case of multiple linear regression, SSE will increase and SST remains constant. This increases the R^2 value even there is no significant relationship between predictors and target. To solve this , we use Adjusted R^2 (I.e.) R^2 is adjusted by normalizing both SSE and SST. Generally adjusted R^2value will be less than or equal to R^2value. No increase in adjusted R^2 value after adding new variable to the model shows that the new variable is insignificant or it is not explaining the variation in response variable.

Considering only the R^2 to finalize the regression model will create an issue. Because a high R^2 value is not necessarily a good indicator of the correctness of the model; it could be of spurious regression.

Spurious Regression / Non-sense Regression:
Spurious means false or fake. Spurious regression happens when there is high R^2 value. Variables without any relationship can have a very high R^2 . For Instance; Imagine that we have variables No. of Twitter users every year (predictor) and No. of people died every year (target). Here the R^2 value will be high since there is no actual relationship between these two variables.

Spurious relationship or correlation is a mathematical relationship in which two or more events or variables are not casually related to each other.

Measures of Prediction accuracy:

- **Root Mean Squared Error (RMSE):** Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells

you how concentrated the data is around the line of best fit. It's the measure of noise in the system. RMSE value should be as low as possible for the model to be the best model.

- **Mean Absolute Percentage Error (MAPE):** It denotes the average of residuals. If the value of MAPE is low for a model, then the model is the best model. Since MAPE is a measure of error, high numbers are bad and low numbers are good.

  (Always the residual should be low for the model to be the best model)

- Residual Analysis is seen in previous pages during assumptions content.

Limitations:

- Linear regression is made on assumption that our data have linear relationship. But, in many real-world scenarios, that's not true.
- Fast and easy to model and is particularly useful when the relationship to be modeled is not extremely complex and if you don't have a lot of data.
- Linear Regression is very sensitive to outliers.

CHAPTER 4

LOGISTIC REGRESSION

Logistic Regression

Types of Logistic regression

Methodology: Logit Transformation

Assumptions of Logistic regression

Issues of assumptions & remedial measures

Measures involved in logistic regression & interpretation

Evaluation techniques

CHAPTER 4

LOGISTIC REGRESSION

When the target variable is categorical (binary), we go with logistic regression. We do regression and classify the outcomes based on based scenarios.

"Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function"

In the mathematical side, the logistic regression model will pass the likelihood occurrences through the **logistic function** to predict the corresponding target class.

For instance; How likely the students get admission in IIT?

How likely the customer entering the shop will make a purchase?

What's the probability that my girl friend will accept my proposal?

Types of Logistic Regression:

Binary logistic regression: When the dependent variable has only two factors (binary), we do binary logistic regression. Eg: Spam/ Not Spam, 0 / 1

Multinomial logistic regression: When the dependent variable has more than two factors, we do multiple logistic regression. Eg: Lower class/ Middle class/ Upper class , Small/ Medium/ Large/ XL.

Ordinal logistic regression: When the dependent variable has more than two factors with ordering. Eg: Movie or Service Ratings from 1 to 5

Methodology of Logistic Regression:

Our objective is to draw the best fitting S or Sigmoid curve to predict the probability of occurrences of target. We do the following procedures for it.

Logit Transformation:

Here what we have on Y axis is probability of target variable. P(Y) = f (independent variables). We find the probability of occurrences of target variable. Probability values are restricted to 0 to 1.

But the regression values will be unbounded (- infinity to + infinity). But in logistic regression, because of probability concept the values will be 0 to 1.

To solve this, we do logit transformation. I.e. log of odds ratio of target. This transformation allows us to come up with linear relationship with predictors, which is also consistent with what is on left hand side of equation which is probability of Y.

Odds ratio: It is the statistical term that denotes probability of success to probability of failure.

$$Odds\ Ratio = \frac{P}{1-P} = \frac{Probability\ of\ Success}{Probability\ of\ Failure}$$

P/1-P can take values of 0 to infinity

Then log (odds ratio). Thus, will take values of $-\infty\ to\ +\infty$. The prediction values from a linear regression variable can have values between $-\infty\ to\ +\infty$and on the left-hand side the log(p/(1-p)) can also take the values of $-\infty\ to\ +\infty$. So, we don't have an inconsistency problem.

The equation becomes

$$\log\left(\frac{P}{1-P}\right) = Y = \beta 0 + \beta 1(X)$$

Now the interpretation can be done as ; a unit change in X will lead to fixed % change in log Y

This value pushes the value to +ve and -ve infinity in the plot. Now we project the values to the line and assign the appropriate log odds value. Then transform the log odds to probabilities by using

$$P = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}}$$

is used to again get the S curve and the maximum likelihood is estimated.

The Coefficients of logistic regression are estimated using a technique called Maximum Likelihood Estimation (MLE).

Now we use the observed points on s curve to calculate the maximum likelihood given the s curve. Then the iterations are done using rotating the linear line and the above procedure is continued.

At last the S curve which gives the maximum value for likelihood is the best curve. Then using this we will be able to predict the target variables and classify them in to levels in target variable.

Assumptions:

Logistic regression doesn't have many intricate assumptions when compared to linear regression. Logistic regression does not require linear relationship between predictors and target, residuals need not to be normally distributed, no need of homoskedasticity.

But Logistic Regression has the following assumptions;

- The target variable should be categorical
- The ordinal logistics regression needs the target variable to be ordinal.
- There should not be outliers in the continuous predictor variables.
- There should not be multicollinearity
- There should be the linear relationship between logit of the outcome and each predictor variables. It does not require linear relationship between target and predictors. But it requires the linearity between predictors and log odds.

Issues and treatments:

Multicollinearity: (as seen in linear regression)

Linearity between log odds and predictors:

How to check linearity? Remove the categorical predictors from the dataset. Now bind the logit values to the data. Then create the scatter plots for logit values and each predictor. We can find that which predictors are linearly related to logit value.

Effect: Otherwise the test underestimates the strength of the relationship and rejects the relationship too easily, that is being not significant (not rejecting the null hypothesis) where it should be significant.

Remedial Measure: A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the model. Another approach would be to use discriminant analysis, if the assumptions of homoscedasticity, multivariate normality, and absence of multicollinearity are met.

We need more info on the distributions of each of these variables and number of cases. One thought is whether transforming your independent variable might yield better results. If the mean value is normal, could you transform it into quartiles and see if you get a different/significant result? Additionally, you could group your sample by another variable in your dataset and see if relationships arise.

Measure involved in Logistic Regression:

p value is the probability value found by assuming null hypothesis is true. Based of p value we accept or reject the null hypothesis. Generally, p value shows how significant the variable in linear regression. Predictors which has p value with star ratings (*** ,**, *) in the model result are considered as the highly significant variables. Other predictors are omitted.

Coefficients:

For every one-unit change in X1, the log odds of Y increase or decrease by certain value. It is effect of one variable on the log odds of target variable. If the predictor variable is categorical then the interpretation of it differs.

Evaluation Metrics of Logistic Regression:

Testing Goodness of fit of model:

Likelihood Ratio Test: In Logistic Regression, likelihood ratio test is used for used for checking the statistical significance of the overall model.

The likelihood-ratio test is a hypothesis test that compares the goodness-of-fit of two models, an unconstrained model with all parameters free, and its corresponding model constrained by the null hypothesis to fewer parameters, to determine which offers a better fit for your sample data.

Hypothesis testing is being done here;

H0: The model is not statistically significant

H1: The model is statistically significant.

Based of p value of LR test, we conclude whether the model is significant or not.

Pseudo R^2: R^2 is used to find the goodness of fit on model in Linear Regression. Generally, R^2 is calculated from residuals. But in Logistic Regression, the residuals are infinite. So, we go with new measure called **Pseudo R^2** .

There are various Pseudo R^2 techniques such as;

Mc Fadden's pseudo R^2

Cox & Snell R^2

Nagelkerke's R^2. And their p values are found to accept or reject null hypothesis.

Validation of accuracy of prediction:

Confusion Matrix: Confusion matrix or error matrix is used to describe the performance of classification algorithm.

	Class 1 Predicted	*Class 2 Predicted*
Class 1 Actual	**TP**	**FN**
Class 2 Actual	**FP**	**TN**

•True Positive (TP): Observation is positive, and is predicted to be positive.
•False Negative (FN): Observation is positive, but is predicted negative.
•True Negative (TN): Observation is negative, and is predicted to be negative.
• False Positive (FP): Observation is negative, but is predicted positive.

Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Kappa: Kappa is used to check accuracy of classification model. Kappa > 60% is the best classification model.

ROC Curve: ROC Curve is used to understand the overall worth of a logistic regression model. It is used as an accuracy measure in classification problems.

Gain Chart: Gain and Lift charts are used to evaluate performance of classification model. They measure how much better one can expect to do with the predictive model comparing without a model. It's a very popular metrics in marketing analytics.

Thank you

Regards,
Ranjith P
Email Id: pgrranjith@gmail.com