Predicting the "Re-admission possibility of a patient into the hospital"

Abstract:

The major drawback in the health care industry when coming to medical facilities is that when a person who was been discharged after treatment and readmits to the hospital with in a stipulated amount of time. This leads to severe image damage of the hospitals.

In order to avoid the readmission of the treated patient with in that time period we will be going to make a model to predict the readmission of the patients. So, that a proper action can be taken in order to take a precautionary action to reduce the patients from readmitting.

Problem Domain Dataset

A leading hospital in the US is suddenly seeing increase in the patient readmission in less than 30 days. This is serious concern for the hospital as it may indicate insufficient treatment or diagnosis when the patient was admitted first and later released under clean bill of health. Not only the image of hospital as healthcare provider is compromised, this is also increased cost to the entire Medicare ecosystem in form of increased insurance claims.

Hence it is in Hospital's interest to support their diagnosis by a better predictive model which you are going to build. Here the objective is: Classify the patients treated by this hospital into two primary categories:

- **Readmitted within 30 days**
- **Not readmitted**

Encounters (Records) As stated:

The dataset contains encounters that satisfied the following criteria:

- It is an inpatient encounter (a hospital admission).
- It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.
- Medications were administered during the encounter.

Features (Attributes):

The attributes represent patient and hospital outcomes. This data set mostly contains nominal attributes such as medical specialty and gender, but also includes a few ordinal attributes such as age and weight and continues attributes such as time(days) in hospital and number of medications.

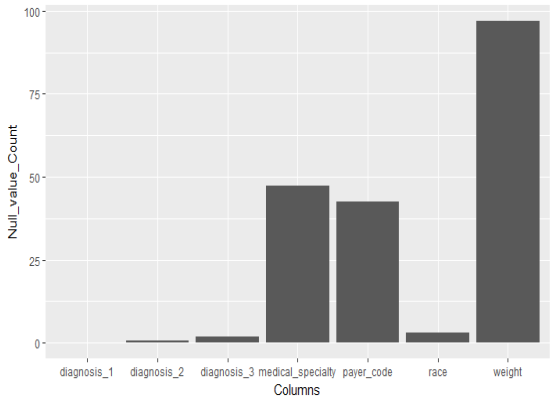

Fig:1 No of "NA" values in the columns

Predicting the "Re-admission possibility of a patient into the hospital"

Attributes distribution:

The total no of different types attributes which are present in the data are:

- Numeric : 7
- Factor : 37
- Target : 1

The Target is binary class only.

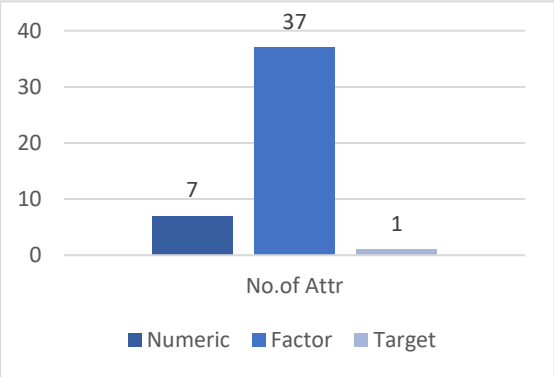

Fig:2 Distribution of variable types

Target Variable:

The last attribute in the previous graph is the class attribute, which in this case is Readmission.

The distribution of the class attribute is as follows:

- Encounters of patients who were not readmitted (No) to the hospital. There are 29,891 of such encounters.
- Other Encounters of patients who were readmitted to the hospital within 30 days of discharge.

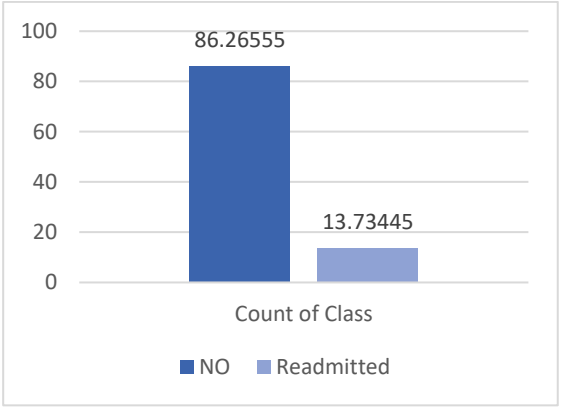

Fig:3 Distribution of Target Class

Data Cleaning Process

Data cleaning is commonly defined as the process of detecting and correcting corrupt or inaccurate records from a dataset, table, or database.

- Data quality is an important component in any data mining efforts. For this reason, many data scientists spend most of their time preparing and cleaning their data before it can be mined for insights.
- There are four broad categories of data quality problems:
    1. missing data
    2. abnormal data (outliers)
    3. departure from models
    4. goodness-of-fit
- For this project we will be using some resampling techniques to handle the imbalance data and even as per the time constraints we will be using central imputation to impute the data.

Predicting the "Re-admission possibility of a patient into the hospital"

Missing Values:

The data has three attributes with the majority of their records missing such as weight (97%), payer code (42%), and medical specialty (43%).

Weight was not properly recorded since this experiment was done prior to the HITECH legislation of the American Reinvestment and Recovery Act in 2009, while payer code was deemed irrelevant by the researchers.

As a result, these 3 attributes were deleted.

There were also 23 attributes that had zero values in 79% to 99% of their records.

Those are medications features such as metformin and other generic medications.

The zero-value indicated that the type of medication was not prescribed to the patient.

As a result, I took three approaches to tackle this problem

1. All these 23 attributes were deleted.
2. These 23 attributes were clubbed into 6 classes based on their molecular structure.
3. Using all the 23 attributes.

Irrelevant Data:

The attribute, discharge disposition, corresponds to 29 distinct values that indicate patients are discharged to home or another hospital, to hospice for terminally ill patients, or indicate that the patients have passed away.

To correctly include only active (alive) patients and not in hospice, we removed records that had Discharge Disposition codes of 11, 19, 20, and 21.

Feature Engineering and Variables Extraction

Feature engineering is the process of using domain knowledge of the data to create **features** that make machine learning algorithms work. **Feature engineering** is fundamental to the application of machine learning and is both difficult and expensive.

As the drug variables which are not prescribed most of the time due to which there is a huge loss of information.

Most of the time the class No is prevailing in all the variables due to which the information which has to be extracted to predict the readmission.

Drugs clubbed:

- **Sulfonylurea:** Chlorpropamide, Glimepiride, Acetohexamide, Glipizide, Glyburide, Tolbutamide, Tolazamide, Glyburide Metformin, Glipizide Metformin.
- **Meglitinides:** Repaglinide, Met glinide.
- **Thiazolidinediones:** Pioglitazone, Rosiglitazone, Troglitazone, Metformin Rosiglitazone, Metformin Pioglitazone.
- **Biguanide:** Metformin, Glyburide Metformin, Glipizide Metformin, Metformin Rosiglitazone, Metformin Pioglitazone.
- **Glucosides:** Acarbose, Miglitol.
- **Insulin:** Insulin.

Predicting the "Re-admission possibility of a patient into the hospital"

We will be just clubbing all this variable under one common molecule name and just break them into two class whether it was prescribed or not.

No of Days Stayed and Month:

This variable was being extracted based on the admission date and discharge date.

Stayed= discharge – admission

Month from the Admission Date.

Releveling Attributes:

We have various attributes which were having huge number of levels. In order to avoid the huge levels, we will be using some techniques to overcome this.

admission_type_id: As this had 8 levels we made it down to 2 levels such as casual and emergency were Emergency, Urgent and Trauma Centre into one class and others too Casual.

discharge_disposition_id: As this had 29 levels and out of which 4 levels were discarded and others were clubbed into 14 levels.

admission_source_id: As this had 26 levels which were clubbed into 12 levels.

Age: As it was in ordinal form in the given data we converted them into numeric by just taking mean into account.

A1Cresult and Ma Gluc serum: Just did a simple change of converting the old levels into normal, abnormal and not tested.

ICD code Levelling:

The three variables which were diagnosis 1, diagnosis 2, diagnosis 3 where above 600 levels are been clubbed into 13 levels based on ICD 9 codes.

Handling Data Imbalance

What is Imbalanced Classification?

Imbalanced classification is a supervised learning problem where one class outnumbers other class by a large proportion. This problem is faced more frequently in binary classification problems than multi-level classification problems.

What are the methods to deal with imbalanced data sets?

The methods are widely known as 'Sampling Methods'. Generally, these methods aim to modify an imbalanced data into balanced distribution using some mechanism. The modification occurs by altering the size of original data set and provide the same proportion of balance.

Below are the methods used to treat imbalanced datasets:

1. Under sampling
2. Oversampling
3. Synthetic Data Generation

1. Under sampling

This method works with majority class. It reduces the number of observations from majority class to make the data set balanced. This method is best to use when the data set is huge and reducing the number of training samples helps to improve run time and storage troubles.

Predicting the "Re-admission possibility of a patient into the hospital"

2. **Oversampling**

This method works with minority class. It replicates the observations from minority class to balance the data. It is also known as *up sampling*.

3. **Synthetic Data Generation**

In simple words, instead of replicating and adding the observations from the minority class, it overcome imbalances by generates artificial data. It is also a type of oversampling technique.

Model Building and Comparison

Based on the data which where been cleaned and further variables extraction models where been created.

Two Main data frames where been created:

1. Data with all variables.
2. Data with drugs dropped.

The models which where been chosen to predict the readmitted class are:

1. Logistic Regression
2. Naïve Bayes
3. Decision Tree
4. Ada Boost
5. Ensemble Averaged

Logistic Regression:

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

To represent binary / categorical outcome, we use dummy variables. You can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

Naïve Bayes:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

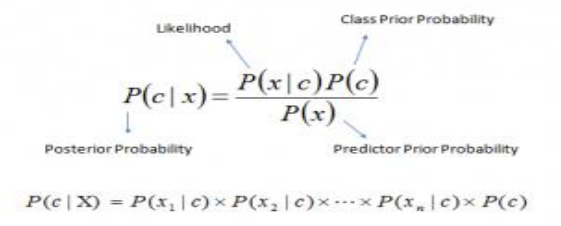

Fig:4 Bayesian Formula

Decision Tree:

Decision tree is a type of supervised learning algorithm (having a pre-defined

target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

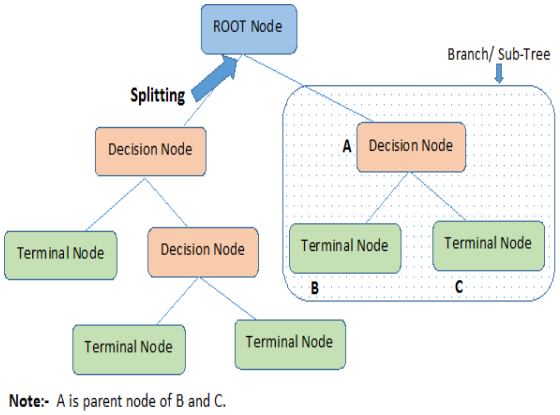

Fig:5 Decision Tree split

Adaboost:

Adaptive Boosting, is a machine learning meta algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.

AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favour of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

Ensembled Average:

This method is an experiment kind of thing as the model which where been used were not giving the desired results as they are weak learners.

So, in order to boost the performance a given datasets majority class is been sampled into n dataset and the minority class is been added to every dataset then N final datasets are made.

Then using that N dataset N classifiers are been made and then prediction and averaged down and the threshold value is being set at maximum AUC on the validation set.

Experiments

Case 1: Using all the variables in the dataset:

As the base line model most of the variables which are having more than 50 levels are been removed and model was build using the decision tree.

By using the feature engineering and data balancing techniques few models were made and evaluated.

Performance of the various model using various techniques:

Predicting the “Re-admission possibility of a patient into the hospital”

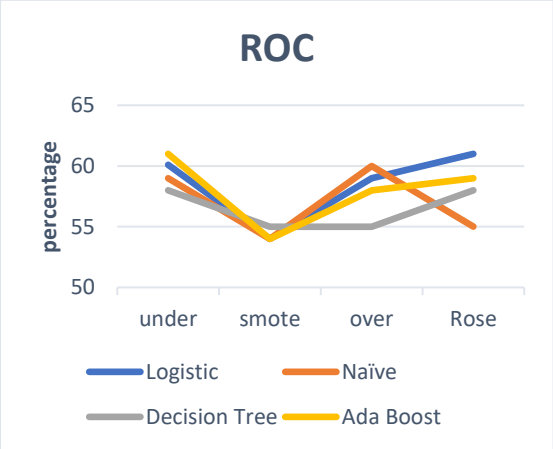

Fig:6 Case 1 ROC

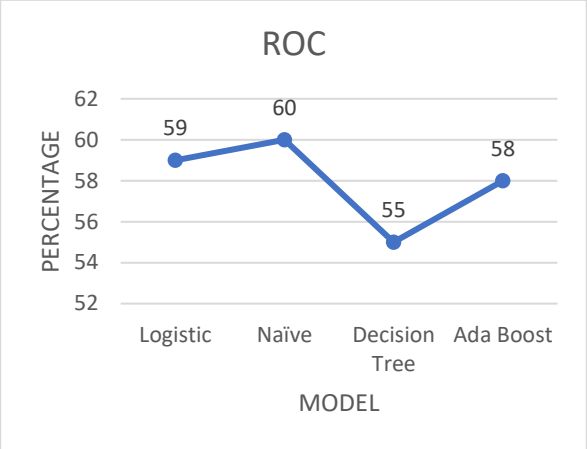

Fig:8 Model Based ROC

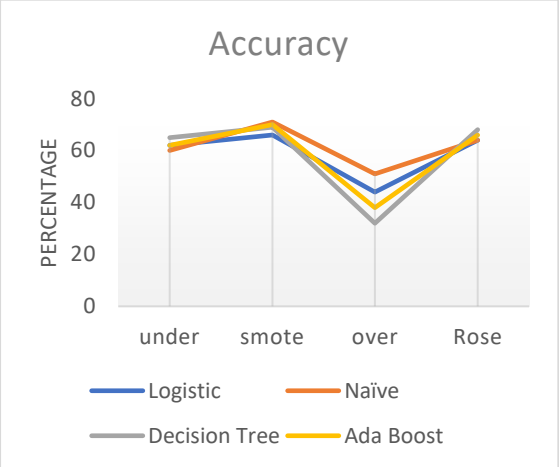

Fig:7 Case 1 Accuracy

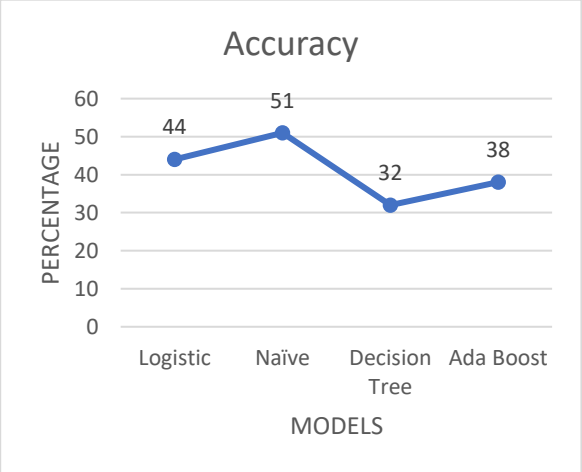

Fig:9 Model Based Accuracy

From this chart above its very clear that over sampling the minority class is working better and giving good consistent results when compared to the others.

Even the under sampling is giving good results but as of now we will not be considering that as it gives tremendous loss of information.

From the above charts it is clear that naïve Bayes is performing well on the dataset with all the variables.

Case 2: Using only feature variables and chi-square test:

As some drugs are not adding up any predictive power to the target variables as most of their class was prevailed by no which means the drug was not prescribed.

So new 6 features where been made and added to the dataset and all other drugs where been discarded.

Predicting the "Re-admission possibility of a patient into the hospital"

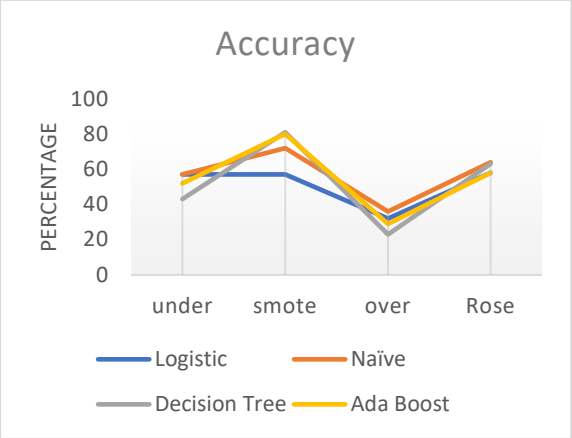

Fig:10 Case 2 Accuracy

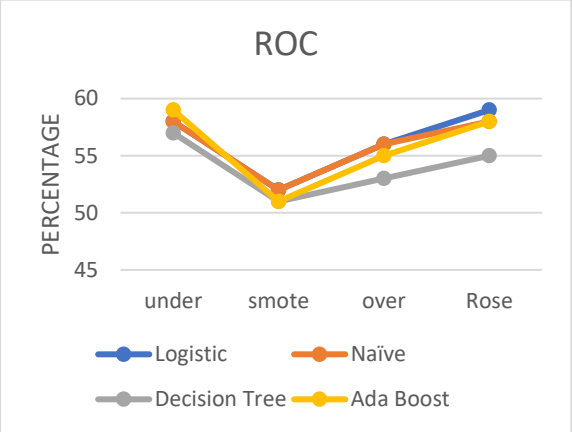

Fig:11 Case 2 ROC

From this chart above its very clear that over sampling by using rose is working better and giving good consistent results when compared to the others.

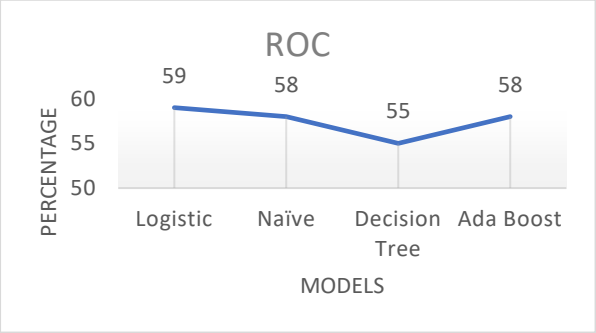

Fig:11 Model Based ROC

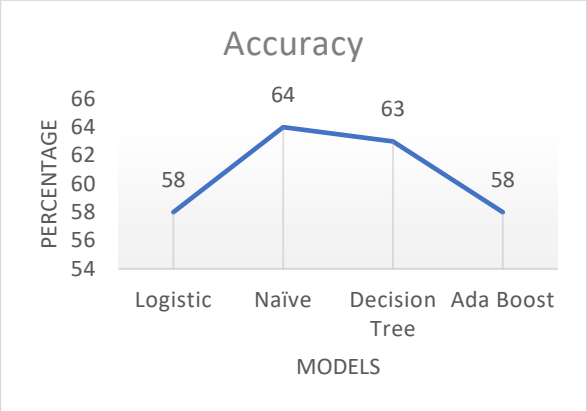

Fig:12 Model Based Accuracy

From the above graph it is clear the Logistic is giving better result the sensitivity in that data is very due to which this model will not be suitable to make as a base to predict the model.

Case 3: Ensemble Technique using the all the variable dataset:

The Models was being designed by using the n sample out of the majority class and minority and making a n model.

This technique is known as boosting as we are combing the week learners to make the strong learner to predict.

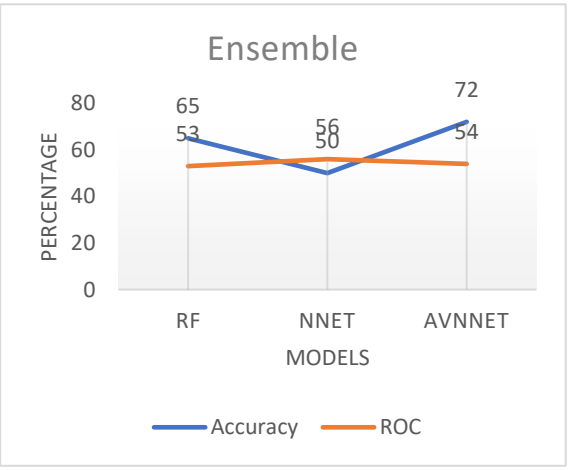

Fig:13 Ensemble model

Predicting the "Re-admission possibility of a patient into the hospital"

Pattern Extraction using the Association Rules Model:

rules	support	confidence	lift	count
{num_diagnoses= [6,9), max_glu_serum=not_tested, discharge_disposition_id=Another_Rehab} => {target=Within30days}	0.010374	0.853448	1.71138	99
{num_diagnoses= [6,9), max_glu_serum=not_tested, discharge_disposition_id=Another_Rehab, glucosides=No} => {target=Within30days}	0.010269	0.852174	1.708824	98
{num_medications= [18,81], discharge_disposition_id=Another_Rehab, insulin=Yes} => {target=Within30days}	0.01006	0.849558	1.703578	96
{num_medications= [18,81], diabetesMed=Yes, discharge_disposition_id=Another_Rehab, insulin=Yes} => {target=Within30days}	0.01006	0.849558	1.703578	96
{num_medications= [18,81], discharge_disposition_id=Another_Rehab, meglitinides_2=No, insulin=Yes} => {target=Within30days}	0.01006	0.849558	1.703578	96
{num_medications= [18,81], discharge_disposition_id=Another_Rehab, glucosides=No, insulin=Yes} => {target=Within30days}	0.01006	0.849558	1.703578	96

Table 1: Association Rules pattern Mining

Summary

As the constraint was to predict the people who will join the hospital with in 30 days so that a proper action can be taken in order to avoid.

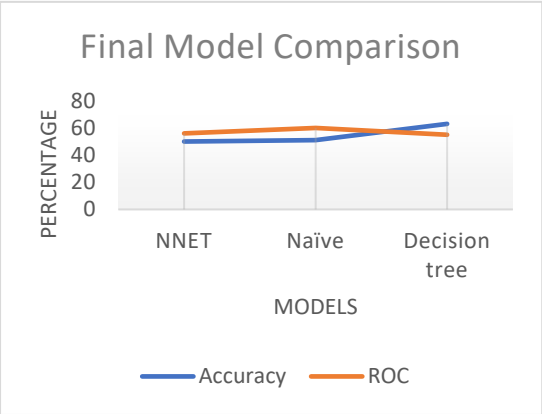

Fig:14 Final model Comparison

As there is strong urge for the interpretability then we will be going for decision tree as ii gives more interpretability when compared to other models.

Even the Naïve Bayes is not being selected as it doesn't capture the linear pattern in the data.

If Metric is important rather than interpretability then we will be going for the Neural net ensemble.

The Variables which are driving the decision the most are:

- Number of days stayed
- Number of Procedures

By Abhilash Reddy for INSOFE

Predicting the “Re-admission possibility of a patient into the hospital”

- Number of medications
- Number of Diagnosis
- Age
- Discharge Disposition ID

References:

Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition, Elsevier, 2011

http://topepo.github.io/caret/train-models-by-tag.html

https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182900

Tamraparni Dasu and Theodore Johnson, *Exploratory Data Mining and Data Quality*, Wiley, 2004.

Wikipedia - https://en.wikipedia.org/wiki/AdaBoost

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

http://www.rdatamining.com/examples/association-rules