Model Accuracy – Transformation of Data to decision.

In Data science, where Data is considered the fuel for moving ahead, the next thing everyone has on the mind, is to know that what value the data is going to bring on the table. The accuracy of any model is moment of truth wherein the organization reap the benefits of collecting, maintaining, storing and processing data. If the accuracy of the model is not as expected, all the endeavor, right from collecting it to arrive at conclusion falls in line.

The blog here is an attempt to outline the ideas that have been useful to enhance the accuracy of the model. There are basically three aspects, which need to be brought to focus to achieve the desired.

Model Accuracy

Data

Algorithms Algorithm Engineering

Tuning

Data Engineering: The first aspect of gaining good accuracy is to condition the data, there lies colossal difference between the data, that is collected at source and the data that is declared fit for model to train. Hence the data needs to be trimmed, conditioned and Engineered, Data Engineering in any model constitutes of following pillars.

Train Data: Having adequate data is always a good idea, it allows the data to tell for itself instead of relying on assumptions and weak correlations presence resulting in inaccurate models. There are situations where we don't get an option to add more data, conditions when we do not get a choice to increase the size of training data, ex. data science competitions, but while working on a company project it is suggested you ask for more data if possible, this will reduce your pain of working on limited datasets technique.

In absence of adequate train data, it is suggested that we divide or segment that data and train the model iteratively. As known that overfit model will tend with inflate the output, the underfitted model will not be able to utilize the algorithm’s capacity to it’s fullest resulting in not so precise results.

Outlier handling: Presence of an outlier values in the training data often reduces the accuracy of a model and creates biases in model, it leads to inaccurate predictions, this is because we don't analyze the behavior in relationship with other variables correctly. Hence it is important to

treat an outlier value. The next point immediately that strikes is how do we detect an outlier for a data set, hence an easy check list to achieve the same is as below

- An outlier is generally defined a value, which is beyond the range of -1.5 x IQR to 1.5 x IQR.
- Define a range i.e. 5% to 95% of all the data to be looked upon as normally distributed and any data falling beyond the same will be treated as an outlier.
- Data points falling out of 3 or 4 standard deviations, depending upon the organizational needs can considered as outliers.

As a next course of action, once we have the correct approach for identifying the outliers, it would be imperative to think about treating them. The below are key aspects of treating the outliers, however as outlier treatment is integral part of Exploratory Data Analysis or (EDA) it should not be considered as one-time process. As there are multiple ways to handle the exception, hence it should be iteratively performed after checking for outputs.

- ***Deletion***: This is the most empirical and easiest method to treat outliers. We simply ignore the records containing the outlier values.
- ***Imputation***: This is method is mostly implied while handling missing values, The idea over here it is replace the outliers with Mode, Mean or Median, so that we do not lose on number of records.
- ***Exception***: This method is to treat when outliers are significant in proportion and cannot be dropped, hence create a separate group of outliers and train the model on the same.

Missing values: Imputation of missing values or “Nan” in data will impact by large on accuracy of model. It is mandate that we check the distribution of data post imputing the missing values to assure that the new values do not create any biases, co-relation in the data set.

Like as discussed in outlier treatment missing values are also to be considered before they are ingress in to algorithm as train data. Missing values when present significantly in train data can lead to underfit the
model. Below are some of the prominent procedures that can be deployed to overcome the problem of missing values.

Imputation: When it comes to missing values or “Nan” in data, the most practical method is to fill the records with Mode, Mean or Median. The benefit of doing the same is that the train data will not be largely effect by the values added. As an added advantage to this method is that if the there were any records that could be have outliers which are now missing are automatically treated.

Prediction Model: The very basic idea of this method, is if the model can used to be forecast from the given data, then it would be pragmatic to predict the value using the

model. This approach splits the data set, one with complete values and other with missing ones. We predict the missing values based upon the train values that were injected to model. The main drawback of the procedure is that if the train dataset has no relationship between the attributes of the missing values, then the result would not be very much precise.

Feature Engineering: This step helps to extract more information from existing data. New information is extracted in terms of new features and these features may have a higher ability to explain the variance in the training data thus giving improved model accuracy. Feature engineering is highly influenced by hypothesis generation, that result in good features that's why it is always suggested to invest quality time in hypothesis generation.

The following methods are to rescue when it comes to Feature Engineering.

PCA: For large dataset, Principal Component analysis (PCA) is the way to go approach, It creates new features based on existing ones.

Dummy variables: When we need to use categorical values are numerical ones, creating Dummy variables is the most widely accepted method.

Feature Selection: Feature selection methods are used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. It is a process of finding out the best subset of attributes which better explains the relationship of independent variables with target. features selection is based on various metrics like domain knowledge and visualization and statistical parameters

Algorithms: After discussing the most prominent part of attaining accuracy the next important aspect is selection of algorithm. Having satisfied all the requirements as mentioned in Data Engineering there is still lot to be done to achieve the accuracy. Algorithm selection takes the second position in the same. Hence below points are needing to be considered while selecting the algorithm.

- The best method is to select the algorithm is to revisit the problem statement again. By having deep understanding of problem statement, we can decide that what algorithm should be implemented to achieve the desired output.
- Another way of selecting an algorithm after sufficing the above-mentioned condition, is to look at the dataset and try to match with the working fundamentals of the algorithm. For eg., If we have multiple classes to identify and our dataset has all the required inputs, it would be viable to go for KNN over Logistic regression.

• The third is to use ensemble learning, where multiple algorithm are brought to compete to get the best accuracy. After fulfilling the above mentioned, this method is the most industry accepted one and is widely used.

Algorithm Tuning: The third and mostly ignored aspect of accuracy improvement is algorithm tuning. Default parameters of algorithm are always in play whenever the mentioned algorithm is brought in action. It would not be out of the place to point out the fact that parameters are also outcome of some learning/training process. Setting the optimum value of parameter in any algorithm can bring the best out it. For eg. In Random Forest (RF), there are host of parameters like *n_estimators, criterion, max_depth* and many more. Just by changing these we have multiple values of accuracy of the model with same train data set.

Accuracy for any model or prediction for any given dataset is any ongoing journey. It is an evolving process until the model is satisfactorily able to add value of the data collected by organization.