**Predicting loan e-signing– Binomial Classification**

1. **Problem Statement:**

A financial company takes the customer data from P2P lending website and asses whether the person will e-sign the Loan

Challenge : Create the classification model to predict the customer will e-sign [1] or will not e-sign [0] based on the various features of the customer like age, income, credit\_score etc

1. **Import Data & Review**
   * **17908 samples, 21 features**
   * **no null values**

* **Numeric 17**
* **Categorical 1**
* **Boolean 3**
  + **53.8% e-signed the loan**

1. **EDA**

* **Histogram results shows good.Most of the features are right skewed. But 'Months\_employed'shows that the data cured for it is not accurate as most of data shows towords 0 - 1.**
* **No abnormal distribution of data in Bi/Multi nomial categorical data**
* **Top 3 postivity correlated feaures with Traget is ['risk\_score', 'amount\_requested', 'has\_debt']**
* **Top 3 Negatively correlated feaures with Traget is ['home\_owner', 'personal\_account\_m', 'age']**
* **Correlation Matrix helps to identify the most correlated features.If 2 features are highly corelated then we can keep only one and remove other to overcome multicollinearity.**
* **In the above results the scale shows from -0.2 to +0.3 which is not too high. So most of the features are not high correlated.**
* **income is correlated with amount\_requested**
* **years\_employed is correlated with current\_address\_year**
* **amount\_requested is correlated with risk**
* **risk 2,3,4 are correlated with each other**

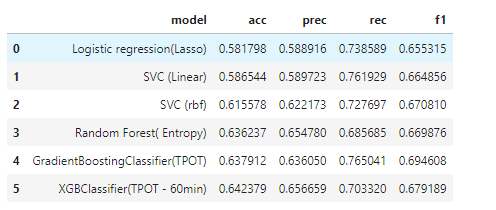
1. **Feature Engineering**
2. **Removed the months\_employed column as this is not accurate**
3. **Combined personal\_account\_m & personal\_account\_y to personal\_account\_months**
4. **Convert categorial variables to numerical by pd\_dummies(remove\_first)**
5. **Saved the userid to a new variables for future evaluation**
6. **Categorized the independent features and target to X & y variables respectively**
7. **Split the data to train & test with stratify on y**
8. **Target values were equally distributed in train & test**
9. **Apply standard scaling. As this loses rows & columns need to take of it.**
10. **Machine Learning model**

**Logistic Regression, SVM, RandomForest were applied. And below are the results**

* **Among them Random Forest Classifier Accuracy of 63% is consistent**
* Precision tells us about the success probability of making a correct positive class classification
* Recall explains how sensitive the model is towards identifying the positive class
* Along with Accuracy, Percision and recall scores are important in assesing the model
* Classification model should have low FP & FN compared to TP

1. **Improving the model by AutoML TPOT**

* **Running the TPOT Classifier for 20 min suggested GradientBoosting Classifier**
* **On re-running for 60 min gives XGBClassifier and gives the accuracy of 64%**



1. **Summary**

* The accuracy of the Logistic Regression and Random Forest gives 58% & 63% respectively.
* **XGBClassifier** was suggested by TPOT AutoML and now the accuracy moved to 64%