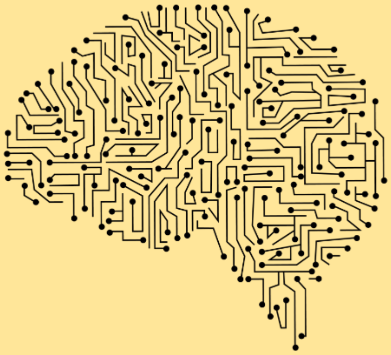
CREDIT RISK ANALYZER

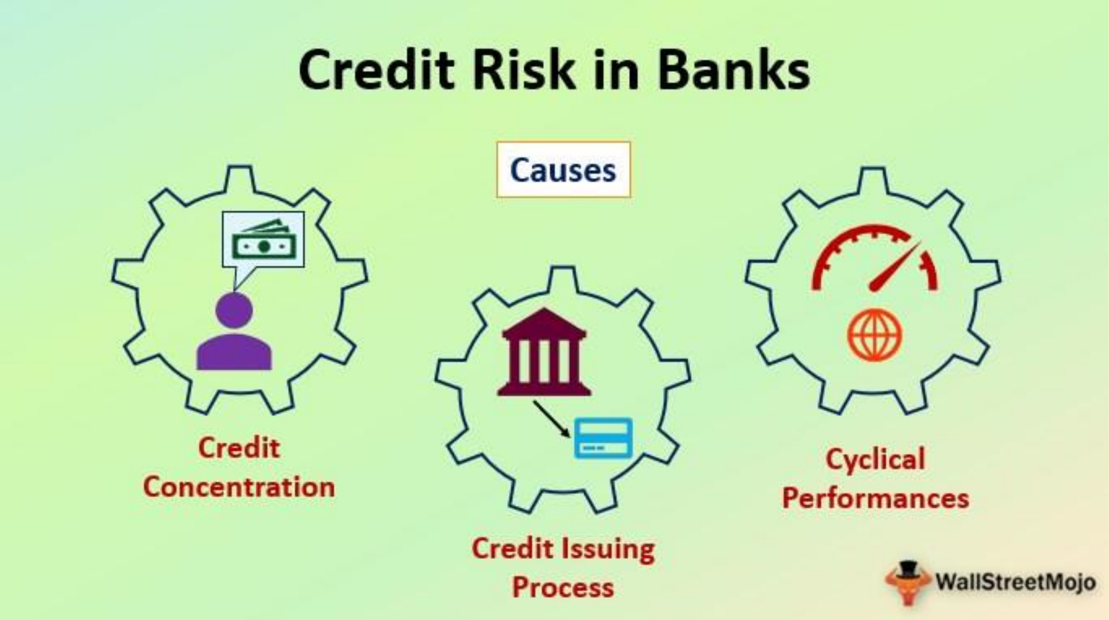


-BY

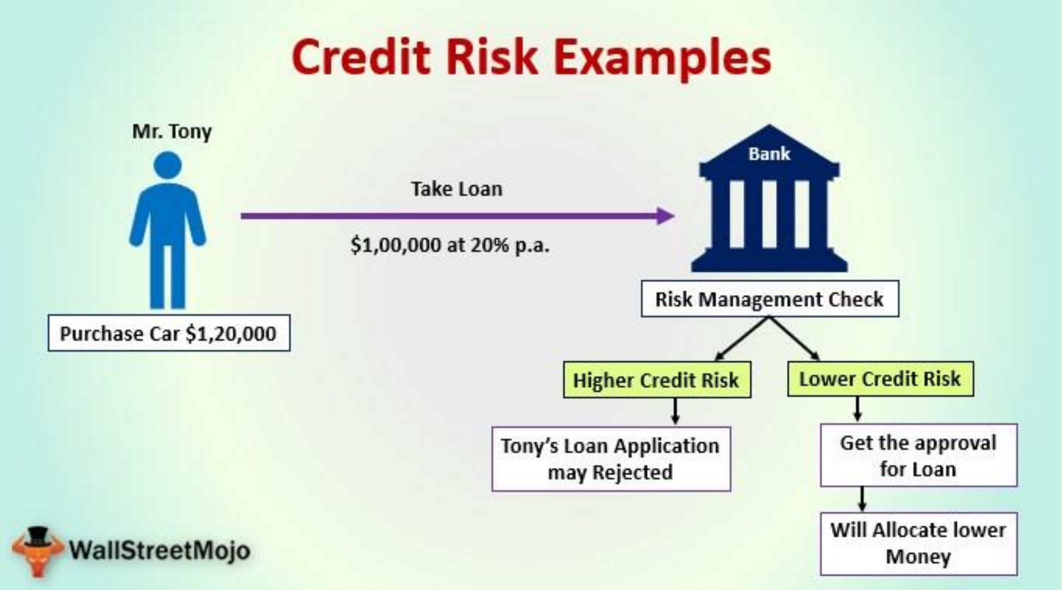
Shivam Kumar Giri

CREDIT RISK

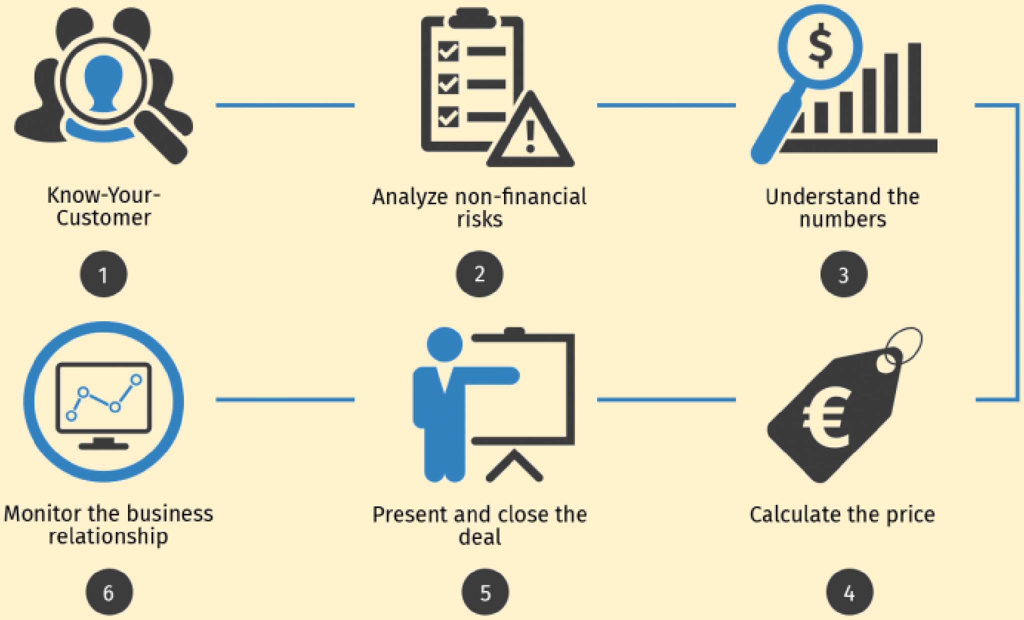
A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments. The risks are calculated on the borrower's ability to repay the loan.



To assess the credit-risk the lenders, look at the five C's of the borrower. The five C's are credit history, capacity to repay, capital, the loans condition, and associated collateral. Some companies have a dedicated department only for assessing the credit risk of its current and potential consumers.



CREDIT RISK WORKFLOW AND CHALLENGES



1 Know-Your-Customer

2 Analyze non-financial risks

3 Understand the numbers

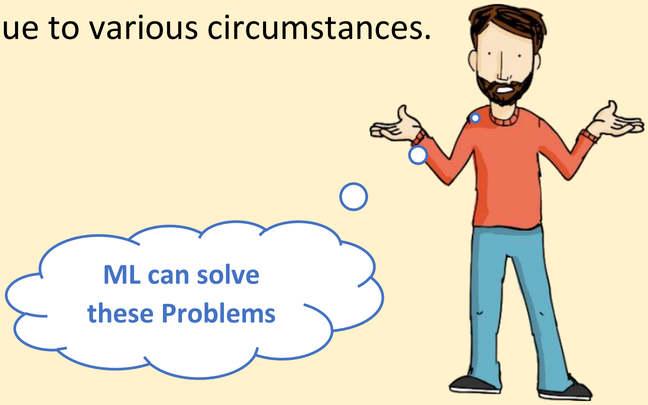
4 Calculate the price

5 Present and close the deal

6 Monitor the business relationship

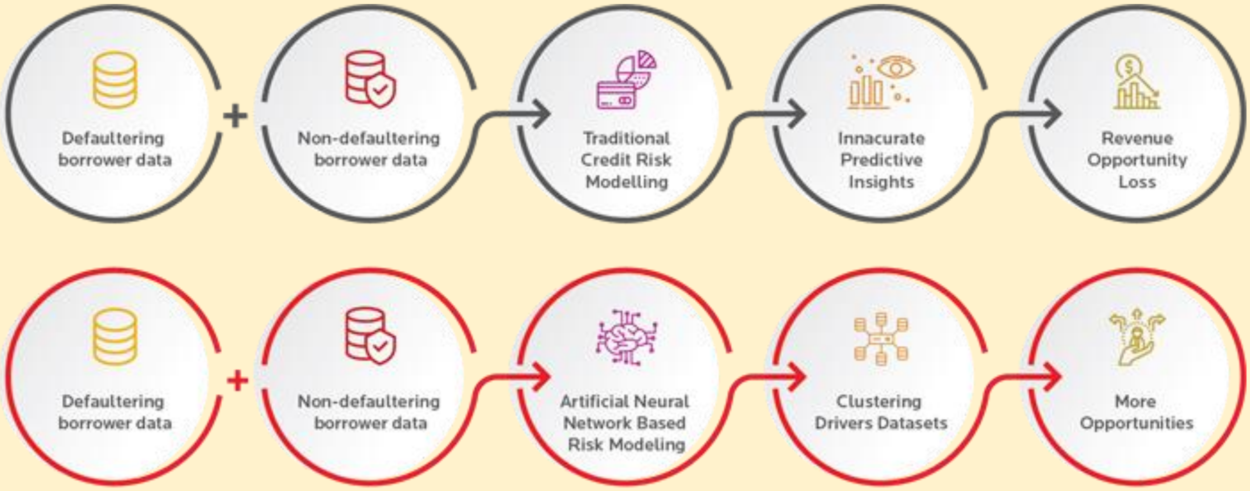
Challenges to Successful Credit Risk Management

- Inefficient data management. An inability to access the right data when it’s needed causes problematic delays.
- No groupwide risk modeling framework. Without it, banks can’t generate complex, meaningful risk measures and get a big picture of groupwide risk.
- Constant rework. Analysts can’t change model parameters easily, which results in too much duplication of effort and negatively affects a bank’s efficiency ratio.
- Insufficient risk tools. Without a robust risk solution, banks can’t identify portfolio concentrations or re-grade portfolios often enough to effectively manage risk.
- Cumbersome reporting. Manual, spreadsheet-based reporting processes overburden analysts and IT.
- Loan not paid by customers due to various circumstances.



CREDIT RISK ANALYZER and MACHINE LEARNING

Credit risk presents a substantial threat to business organizations. As such, risk managers understand the importance of identifying and quantifying the various sources of credit risk. A key component of this analysis consists of building a full understanding of your customers using machine learning methods which can predict and forecast the future activities of customer. Credit risk analysis of your customers and prospects helps **mitigate the risk of default and nonpayment**.


Defaultering borrower data
+
Non-defaultering borrower data
Traditional Credit Risk Modelling
Innacurate Predictive Insights
Revenue Opportunity Loss
Defaultering borrower data
+
Non-defaultering borrower data
Artificial Neural Network Based Risk Modeling
Clustering Drivers Datasets
More Opportunities

How machine learning models lead to better revenue opportunities

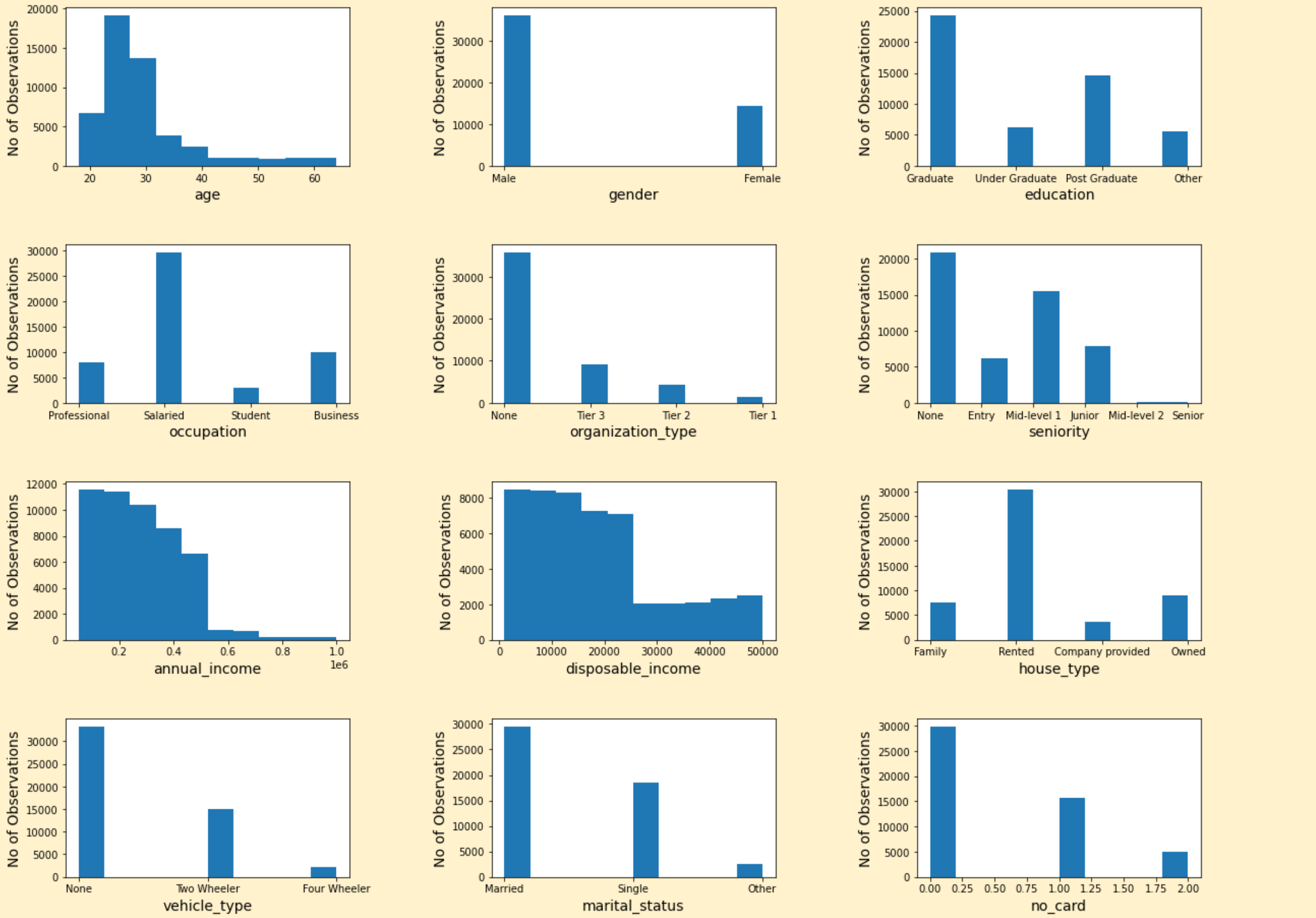
SCOPE OF PROJECT & DATA INFORMATION

Build your custom Credit Risk Analyzer, using a decision tree model, which will **predict the suspected credit card defaulters** among the new applicants. There are 50636 entries in csv file

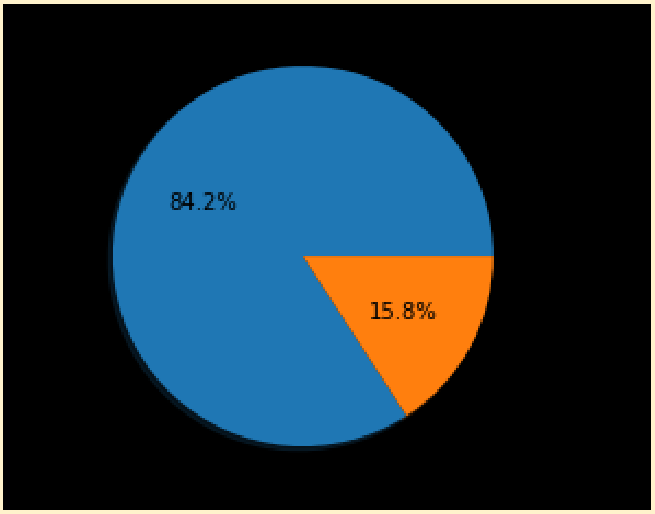
1. **gender** is the gender of the card holder
2. **age** is the age of the card holder.
3. **education** is the last acquired educational qualification of the card holder.
4. **occupation** can be salaried, or self-employed or business etc.
5. **organization_type** can be tire 1, 2, 3 etc.
6. **seniority** denotes at which career level the card holder is in.
7. **annual_income** is the gross annual income of the card holder.
8. **disposable_income** is annual income - recurring expenses.
9. **house_type** is owned or rented or company provided etc.
10. **vehicle_type** is 4-wheeler or two-wheeler or none.
11. **marital_status** is of the card holder.
12. **no_card** has the information of the number of other credit cards that the card holder already holds
13. **defaulter** indicates whether the card holder was a defaulter or not. It is 1 if the card holder was a defaulter, 0 otherwise.

METHODOLOGY

Basic Explanatory Analysis based on various factors of dependence



Also the defaulters ratio is given below there are 15.8% defaulters amounting to 8022 customers:

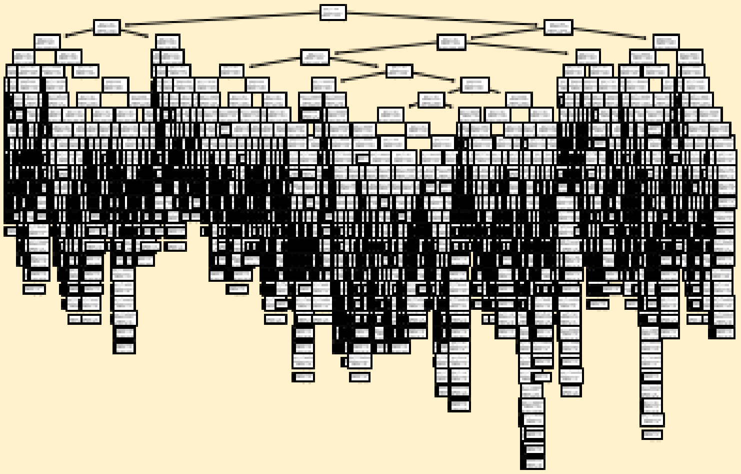


VALIDATION OF DATA AND DECISION TREE

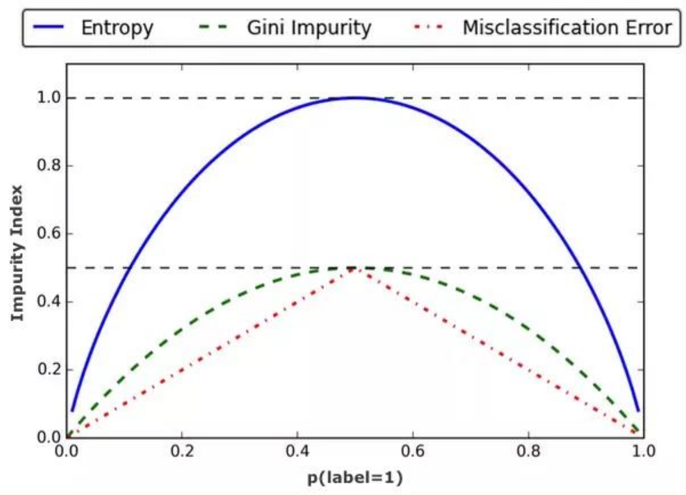
I have basically used 3 methods of classification over two the major factors of optimization functions: gini and entropy:

- Train_test method
- K-fold Method
- Shuffl split Method

Since the size of data is large with 12 factors and 50636 columns the decision tree is explicitly large to be shown, hence very congested.



Hence, I made only one decision tree just for implementation just over entropy method in Train-Test method.

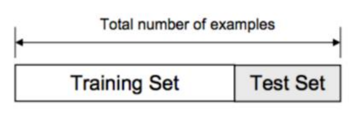


Gini measurement is the probability of a random sample being classified incorrectly if we randomly pick a label according to the distribution in a branch. **Entropy** is a measurement of information (or rather lack thereof), which calculate the information gain by making a split.

TRAIN-TEST METHOD OF VALIDATION OF DATA

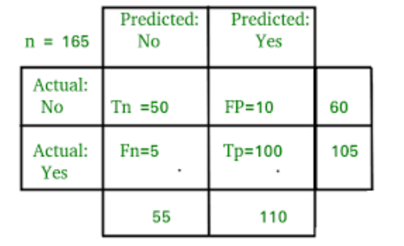
Here Data is splited into set of training data and test data, the training data is trained by decision tree while test data is validated for accuracy.

Here We used 70% of data for training ad rest 30% for testing, which is recommended ratio of splitting.



We used both entropy and gini decision tree and tested their accuracy and the confusion matrix.

Confusion Matrix



n = 165	Predicted: No	Predicted: Yes	
Actual: No	Tn =50	FP=10	60
Actual: Yes	Fn=5	Tp=100	105
	55	110	

Suppose we have total 165 entries with Actual v/s Predicted Entries. Principle Diagonal Elements represent the data accurately predicted where others represent data misclassified.

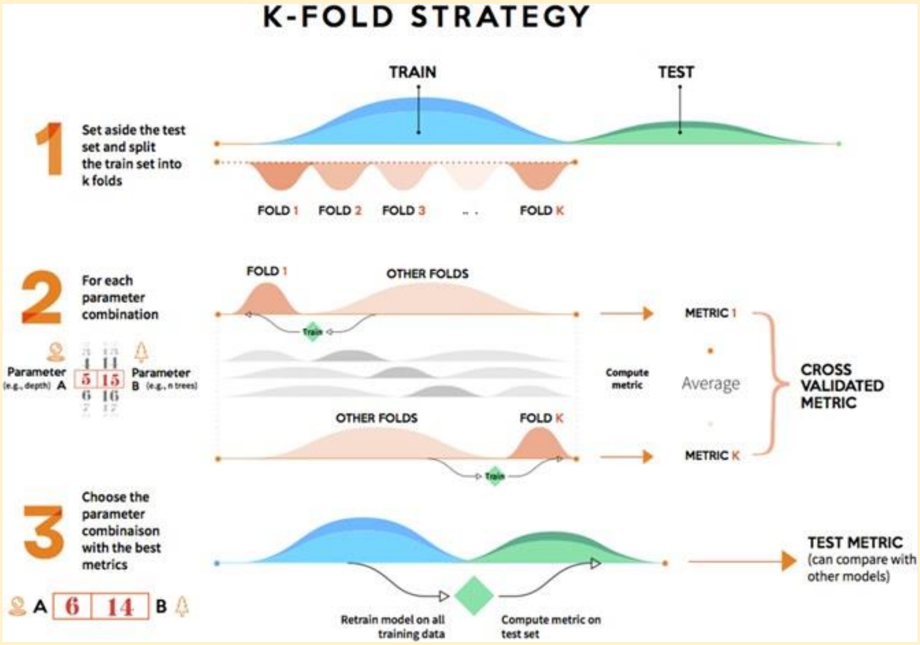
Hypothetical Ideal Confusion Matrix has Accuracies 100%:

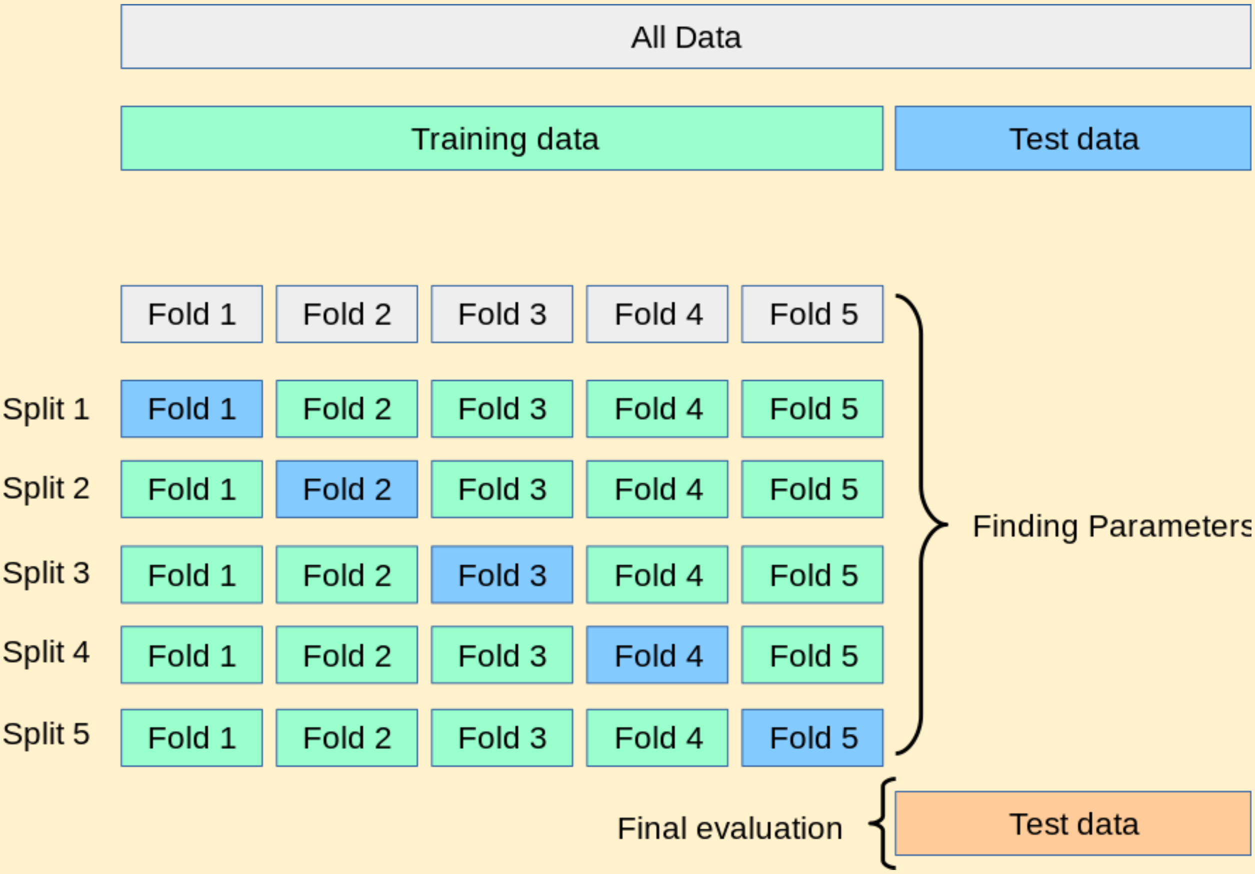
[[100,0]

[0,65]] //Ideal Confusion Matrix with wrongly predicted value 0;

K-FOLD METHOD OF VALIDATION OF DATA

This technique involves randomly dividing the dataset into k groups or folds of approximately equal size. Each fold is then used once as a validation while the k - 1 remaining folds form the training set.


K-FOLD STRATEGY
TRAIN
TEST
1
Set aside the test set and split the train set into k folds
FOLD 1
FOLD 2
FOLD 3
...
FOLD K
2
For each parameter combination
Parameter (e.g., depth) A
5
15
6
16
Parameter B (e.g., n trees)
FOLD 1
OTHER FOLDS
Train
METRIC 1
Compute metric
Average
OTHER FOLDS
FOLD K
Train
METRIC K
CROSS VALIDATED METRIC
3
Choose the parameter combinaison with the best metrics
A
6
14
B
Retrain model on all training data
Compute metric on test set
TEST METRIC (can compare with other models)



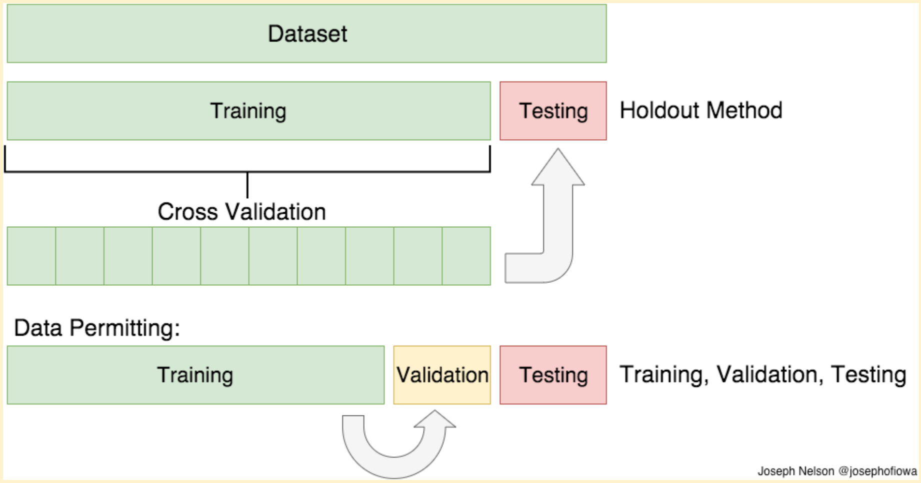
Here we use k=4 ideal for dataset having 30% testing data as per the formula:

K=N/(N*(0.3)) =0.3333, hence k=4, give maximum efficiency.

We used both entropy and Gini decision tree and tested their accuracy and the confusion matrix.

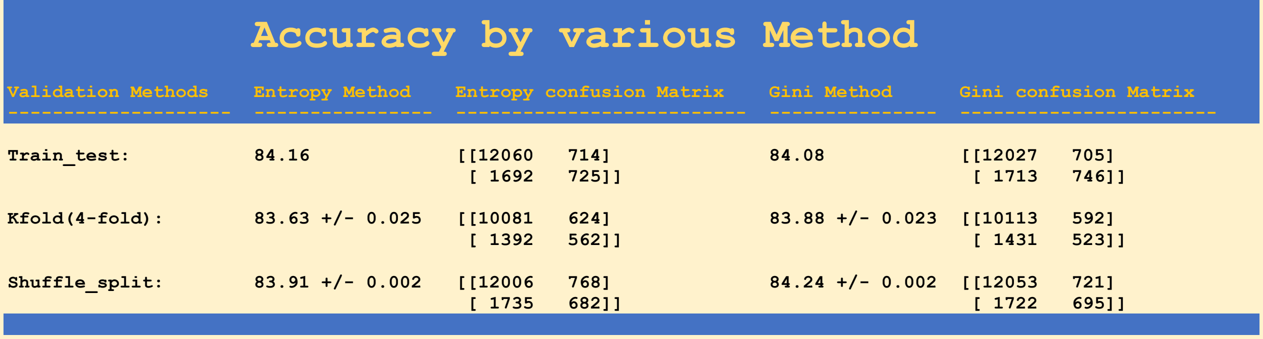
SHUFFLE-SPLIT METHOD OF VALIDATION OF DATA

Shuffle-Split is a Random permutation cross-validator. It Yields indices to split data into training and test sets. We used both entropy and Gini decision tree and tested their accuracy and the confusion matrix.



Joseph Nelson @josephofiowa

OBSERVATION



Accuracy by various Method

Validation Methods	Entropy Method	Entropy confusion Matrix	Gini Method	Gini confusion Matrix
Train_test:	84.16	[[12060 714] [1692 725]]	84.08	[[12027 705] [1713 746]]
Kfold(4-fold):	83.63 +/- 0.025	[[10081 624] [1392 562]]	83.88 +/- 0.023	[[10113 592] [1431 523]]
Shuffle_split:	83.91 +/- 0.002	[[12006 768] [1735 682]]	84.24 +/- 0.002	[[12053 721] [1722 695]]

Hence If we can conclude that we have found maximum Accuracy in case **Shuffle-split Validation method with Gini Optimization method**. It has least deviation of 0.002 from the accuracy.

CONCLUSION

Credit risk analysis is assessing the possibility of the borrower's repayment failure and the loss caused to the financer when the borrower does not for any reason repay the contractual loan obligations. Interest for credit-risk assumption forms the earnings and rewards from such debt-obligations and risks. Our model has successfully analyzed the model and predicted the value with great Accuracies.

Accuracies can be improved with increase of size of data for training as well as testing. Hence our credit risk analyser is ready and can predict the output when required of any new customer with the given factors with a accuracy of 84.24, enhancing the power of ML.