PYTHON CREDIT RISK ANALYTICS – PD MODELS

BACKGROUND:

Credit risk refers to the potential for loss due to failure of a borrower to make a payment when it is due. The risk is mainly for the lender and it can include complete or partial loss of principal, loss of interest, and disruption of cash flow.

The application of scoring models in today's business environment covers a wide range of objectives. The original task of estimating the risk of default has been augmented by credit scoring models to include other aspects of credit risk management: at the pre-application stage (identification of potential applicants), at the application stage (identification of acceptable applicants), and at the performance stage (identification of possible behaviour of current customers)

ABSTRACT:

One of the leading banks would like to predict customers who are most likely to default on the loan.
The sample dataset summarizes the existing customer and new customer details has been provided.

For new customers decide whether to extend credit or not.

DATA SUMMARY:

The data is provided in the csv format, need to be imported using pandas library.

The data contains the credit details about credit borrowers:

age - Age of Customer

ed - Education level of customer

employ: Tenure with current employer (in years)

address: Number of years in same address

income: Customer Income

debtinc: Debt to income ratio

creddebt: Credit to Debt ratio

othdebt: Other debts

default: Customer defaulted in the past (1= defaulted, 0=Never defaulted)

TECHNOLOGICAL REQUIREMENTS:

The following list summarizes the packages used in this project

- Anaconda V – 5.2.0 (py 36_3)
- Python V – 3.6.5
- Packages (Packages that are not part of anaconda distribution) o export_graphviz
  o pydot graphviz o
  pydotplus

METHODS SUMMARY:

Table shows the list of data pre-processing, analysis, visualization and model building techniques applied to complete the project

Task	Task Details	Analytical Techniques	Visualization Techniques
Data Manipulation & Preparation	1. Perform required data manipulation and cleaning. 2. Perform Uni variate and Bi variate analysis	1. Descriptive statistics and outlier analysis 2. Independent t-test and VIF check to get important features	1. Bar Plot, Seaborn (Heatmap), boxplots 2. Boxplot segmentation
Model Building & Performance Check	Create Model and Asses the performance of the models	Build Logistic and Decision tree models, used hyper parameter to fine tune the model.	Visualize the decision tree using graphviz package

MODEL TO PREDICT DEFAULT CUSTOMERS:

- **PART A: DATA MANIPULATION AND CLEANING**

  - Read the input dataset into pandas dataframe –

Load Dataset

```
bankloans = pd.read_csv("Data/bankloans.csv")
```

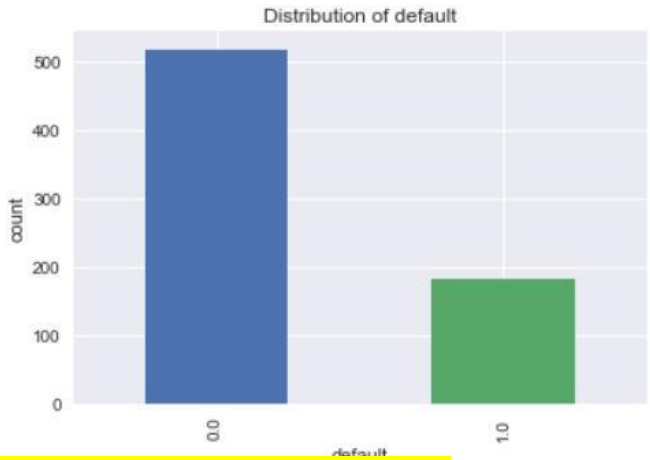
  - Checked for missing values in the dataset

```
#check for number of missing values
bankloans.isnull().sum()
```

```
age          0
ed           0
employ       0
address      0
income       0
debtinc      0
creddebt     0
othdebt      0
default    150
dtype: int64
```

  - There are few missing values in the dataset, on further analysis it is found that missing values in the default column belongs to a new set of customers.
  - Used the concept of winsorization to handle the outliers in the dataset.

Checked the distribution of default and non-defaulted customers in the dataset, to check whether the dataset is imbalanced or balanced data set.



PART B: VARIABLE IMPORTANCE – T TEST AND VIF

- Performed Independent T Test on each variable with 95% confidence level and found that all the variables are with in-significance level.

	Variable Name	T-Statistic	P-Value
0	default	inf	0
1	debtinc	9.95554	3.85688e-20
2	employ	-9.03873	7.73766e-18
3	creddebt	5.20625	3.90256e-07
4	address	-4.82342	2.07201e-06
5	age	-3.83057	0.0001557
6	income	-3.51615	0.000495008
7	othdebt	3.13998	0.00186546
8	ed	3.02788	0.00267847

	VIF Factor	Features
0	41.554332	Intercept
1	1.549227	address
2	2.069008	age
3	2.928049	creddebt
4	5.049334	debtinc
5	1.292872	ed
6	2.624197	employ
7	5.908874	income
8	5.340459	othdebt

- Applied VIF test to find the multicollinearity between the variables, VIF Factor for these variables seems to be with in acceptance levels.

PART C: MODEL BUILDING AND MODEL VALIDATION

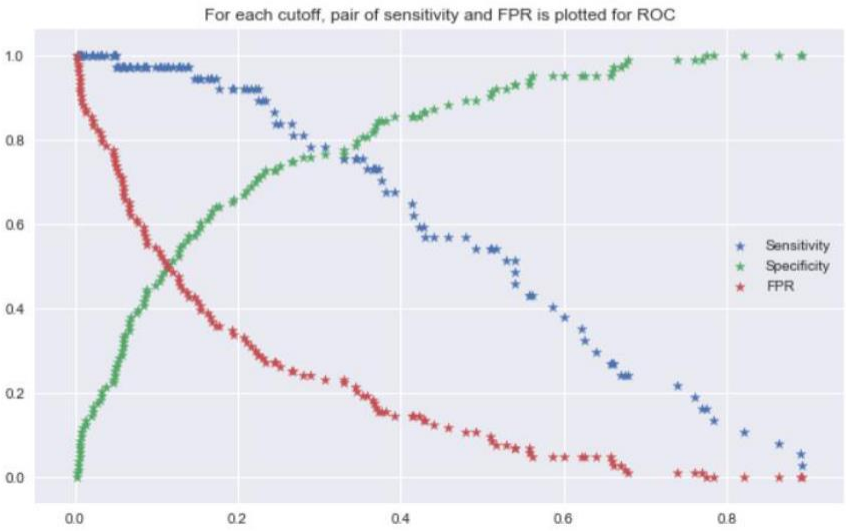
LOGISTIC REGRESSION

- Used stratified sampling based on default, to split the data set into train and test.
- Build the logistic regression model with all the variables and default probability for cut-off is taken as 0.5
- The overall accuracy of the default model is around 80% and recall score (ability of the model to find all the positive samples - find all the default customers) is 54%.
- Even though the accuracy is 80% but it is not a good measure. There are lot of cases which are default and the model has predicted them as not default. The objective of the model is to identify the customers who will default. In this case we need to find the optimum cut-off value

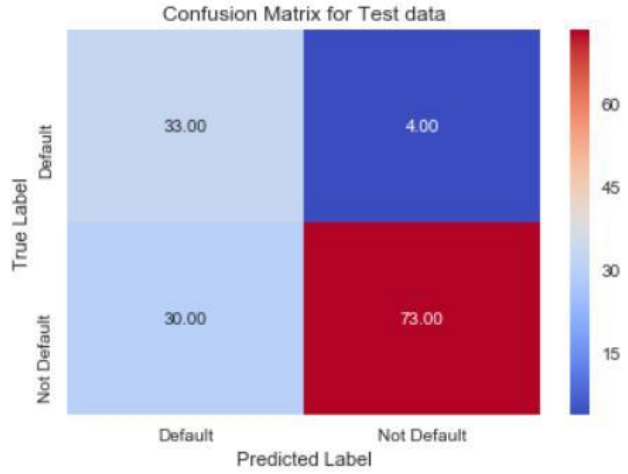
- Found the optimum cutoff value where the sensitivity and specificity is maximum.

```
roc_like_df[roc_like_df['total']==roc_like_df['total'].max()]
```

	falsepositiverate	sensitivity	specificity	cutoff	total
63	0.291262	0.918919	0.708738	0.224326	1.627657



- Created confusion matrix using this cut-off instead of sklearn default cutoff.



- Even though the overall accuracy of the model is reduced from 80% to 75% by taking optimum cutoff as 0.224, Model performance i.e. recall score (ability of the model to find all the positive samples - find all the default customers) has increased from 54% to 89%. The drawback of changing the cutoff value can be seen in drastic drop of precision score (ability of model not to label non default customers as default customers) from 67% to 52%.
- We have a choice to make depending on the value we place on the true positives and our tolerance for false positives, in practical the cutoff values depends on the business decision values.

DECISION TREE CLASSIFIER

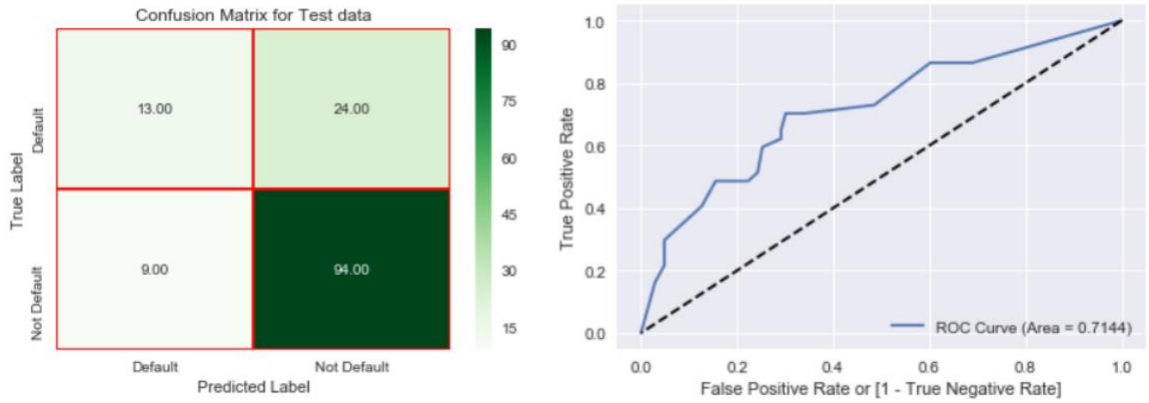
Built the decision tree classifier with 5 fold cross validation, tuned the model using hyperparameters found the best parameters using grid search cv.

```
#display the best parameters for decision tree model

dtclass_model.best_params_
```

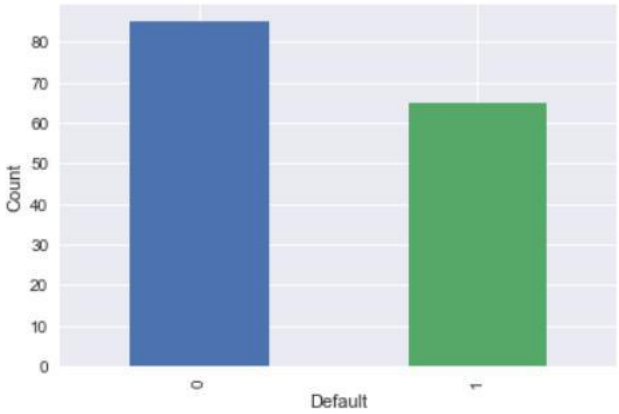
```
{'decisiontreeclassifier__max_depth': 8,
 'decisiontreeclassifier__max_features': 3,
 'decisiontreeclassifier__min_samples_leaf': 2,
 'decisiontreeclassifier__min_samples_split': 14}
```

Confusion Matrix and auc-roc curve for the model,



Model Selection and Business Insights

- Based on the F1-score (harmonic mean of precision and recall), logistic model with f1 score (for positive labels - default customers) of 0.66 is giving better results than decision tree model with f1 score of 0.44. So we will use the logistic regression model to predict the credit worthiness of the customers.
- We will Predict the credit risk for remaining 150 customers using the logistic model with cutoff as 0.224



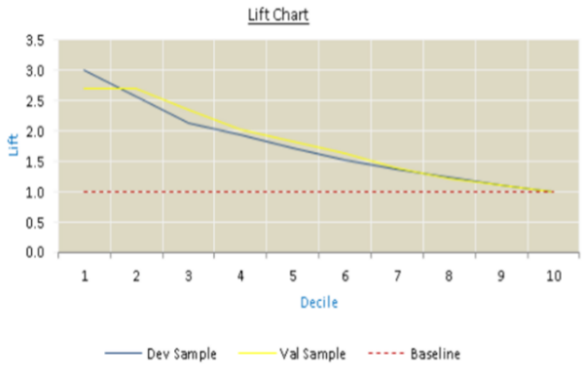
- Out of 150 new customers, model has predicted that 85 customers are not going to default on the bank loan and remaining 65 customers would most likely default on the loan

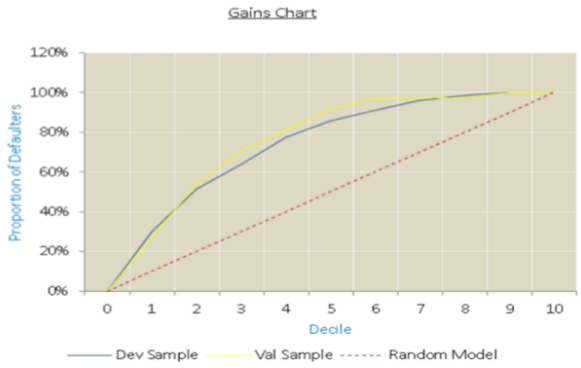
MODEL PERFORMANCE VALIDATION

- KS Chart, Lift and Gain Chart
- we will use the concept of decile analysis for these validations

```
#For Test data set
Image(filename="Images/KS-Testdata.png",width=600)
```

Test Sample									
Decile	Defaulters	Non Defaulters	Total	Default RATE	Default PERCENTAGE	CUMU. Default PERCENT	Non Default PERCENT	CUMU. Non Default PERCENT	KS
1	10	4	14	71.43%	27.03%	27.03%	3.88%	3.88%	23.14%
2	10	4	14	71.43%	27.03%	54.05%	3.88%	7.77%	46.29%
3	6	8	14	42.86%	16.22%	70.27%	7.77%	15.53%	54.74%
4	4	10	14	28.57%	10.81%	81.08%	9.71%	25.24%	55.84%
5	4	10	14	28.57%	10.81%	91.89%	9.71%	34.95%	56.94%
6	2	12	14	14.29%	5.41%	97.30%	11.65%	46.60%	50.70%
7	0	14	14	0.00%	0.00%	97.30%	13.59%	60.19%	37.10%
8	0	14	14	0.00%	0.00%	97.30%	13.59%	73.79%	23.51%
9	1	13	14	7.14%	2.70%	100.00%	12.62%	86.41%	13.59%
10	0	14	14	0.00%	0.00%	100.00%	13.59%	100.00%	0.00%
	37	**103**	**140**					KS	56.94%





- Gain chart tells % of targets (events) covered at a given decile level. In the current case, we can say that we can identify 90% of the defaulters who are likely to default on the loan by just analysing 50% of the total customers.

- Lift chart measures how much better one can expect to do with the predictive model comparing without a model. In the current model, cumulative lift for top two deciles is 2.7, means that by selecting 20% of the records based on the model. One can expect 2.7 times the total number of defaulters to be found than the randomly selecting 20% of the data without a model.