**Deep Analytics and visualization**

**Data science and python**

**creditone project**

Alert! Analytics

**By amarendra ghanekar**

**April 23, 2018**

1 bigData STreet, Ottawa, ON Canada

Table of Contents

[Project Background 2](#_Toc512700168)

[Cleaning and Pre-processing 3](#_Toc512700169)

[Exploratory Data Analysis (EDA) 3](#_Toc512700170)

[Feature Engineering 3](#_Toc512700171)

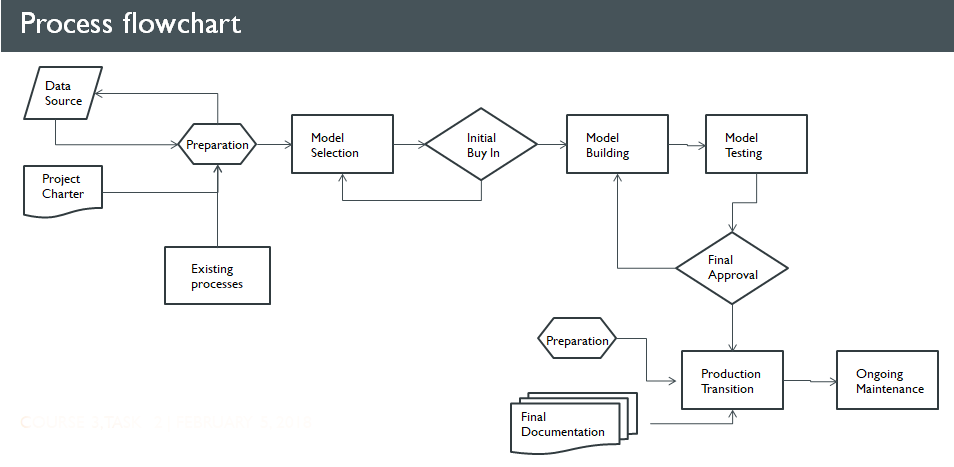
[Classification, model tuning and model selection: 5](#_Toc512700172)

[Observations and Recommendations 7](#_Toc512700173)

# Project Background

CreditOne is credit scoring service and its customers are increasing defaulting loans in last one year. There is a high risk of losing business and hence the problem needs to be addressed immediately. Increase in customer default rates is certainly negative for Credit One, since Credit One approves the customers for loans in the first place.

I followed the process flow as given in the diagram below:



While working on this project, we followed **CRISP-DM (**CRoss-Industry Standard Process for Data Mining), which consists of six steps or phases, as illustrated below:

1.Business (Org) Understanding

Data

6. Deployment 1

5. Evaluation 1

4. Modeling 1

3. Data Preparation

2. Data Understanding

CRISP-DM Conceptual Model.

(Reference: Data Mining for the Masses by Dr. Matthew North)

# Cleaning and [Pre-processing](http://scikit-learn.org/stable/modules/preprocessing.html#preprocessing)

I checked for nulls and other anomalies in the data. Discretization needs of the data, such as the age feature, were explored during this phase. Renaming columns and other data cleanup tasks were completed during this phase.

# Exploratory Data Analysis (EDA)

(Ref: AmarendraGhanekarC5T2.ipynb)



**Some of the key observations:**

**Age:**

Age range is 21 to 79 with mean of 35.

Age can be discretized in 4 buckets

**Limit Balance:**

Around 66% people have Limit Balance of 200,000

Third Quartile or Upper quartile (UQ) - 75% people have Limit Balance of 240,000.

# Feature Engineering

I used RandomForestClassifier () to rank features based on importance

(Ref: AmarendraGhanekarC5T3Variables.ipynb)

|  |
| --- |
| **Feature Importance** |
| PAY\_0 0.083481 |
| PAY\_2 0.073840 |
| ID 0.067933 |
| BILL\_AMT1 0.057422 |
| AGE 0.056730 |
| LIMIT\_BAL 0.053426 |
| BILL\_AMT3 0.048846 |
| BILL\_AMT4 0.048631 |
| BILL\_AMT2 0.048240 |
| PAY\_AMT1 0.048179 |
| BILL\_AMT5 0.046188 |
| BILL\_AMT6 0.044828 |
| PAY\_AMT6 0.042336 |
| PAY\_AMT3 0.041603 |
| PAY\_AMT2 0.041359 |
| PAY\_AMT5 0.039737 |
| PAY\_AMT4 0.037882 |
| PAY\_4 0.021046 |
| PAY\_6 0.019728 |
| PAY\_5 0.019251 |
| EDUCATION 0.018616 |
| PAY\_3 0.016635 |
| SEX 0.012032 |
| MARRIAGE 0.012027 |

I also used correlation and covariation to understand the importance of features in modeling.

(Ref: AmarendraGhanekarC5T2.ipynb)

|  |
| --- |
| **Correlation - default payment next month** |
| ID -0.013952 |
| LIMIT\_BAL -0.153520 |
| SEX -0.039961 |
| EDUCATION 0.028006 |
| MARRIAGE -0.024339 |
| AGE 0.013890 |
| PAY\_0 0.324794 |
| PAY\_2 0.263551 |
| PAY\_3 0.235253 |
| PAY\_4 0.216614 |
| PAY\_5 0.204149 |
| PAY\_6 0.186866 |
| BILL\_AMT1 -0.019644 |
| BILL\_AMT2 -0.014193 |
| BILL\_AMT3 -0.014076 |
| BILL\_AMT4 -0.010156 |
| BILL\_AMT5 -0.006760 |
| BILL\_AMT6 -0.005372 |
| PAY\_AMT1 -0.072929 |
| PAY\_AMT2 -0.058579 |
| PAY\_AMT3 -0.056250 |
| PAY\_AMT4 -0.056827 |
| PAY\_AMT5 -0.055124 |
| PAY\_AMT6 -0.053183 |
| default payment next month 1.000000 |

|  |
| --- |
| **Covariance - default payment next month** |
| ID -50.151705 |
| LIMIT\_BAL -8267.551759 |
| SEX -0.008113 |
| EDUCATION 0.009187 |
| MARRIAGE -0.005273 |
| AGE 0.053143 |
| PAY\_0 0.151499 |
| PAY\_2 0.130960 |
| PAY\_3 0.116867 |
| PAY\_4 0.105115 |
| PAY\_5 0.096020 |
| PAY\_6 0.089194 |
| BILL\_AMT1 -600.394108 |
| BILL\_AMT2 -419.289137 |
| BILL\_AMT3 -405.153680 |
| BILL\_AMT4 -271.199885 |
| BILL\_AMT5 -170.597447 |
| BILL\_AMT6 -132.796294 |
| PAY\_AMT1 -501.374552 |
| PAY\_AMT2 -560.210740 |
| PAY\_AMT3 -411.076284 |
| PAY\_AMT4 -369.515887 |
| PAY\_AMT5 -349.562530 |
| PAY\_AMT6 -392.426415 |
| default payment next month 0.172276 |

* Correlation is a measure that determines the degree to which two variables' movements are associated
* Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated. A positive covariance means that asset returns move together, while a negative covariance means returns move inversely.
* Default payment next moth has positive covariance with PAY\_0 through PAY\_6 and education. Default payment next month has negative covariance with rest of the features. Default payment next month has strong negative covariance with LIMIT\_BAL.

# Classification, model tuning and model selection:

This is a classification problem, and I tested following classification models:

1. **k-nearest neighbors** - KNeighborsClassifier() - AmarendraGhanekarC5T3NN.ipynb
2. **Random Forest** – RandomForestClassifier ()- AmarendraGhanekarC5T3RF.ipynb
3. **Support Vector Machines** - svm.SVC() - AmarendraGhanekarC5T3SVM.ipynb
4. **Decision Tree** – tree.DecisionTreeClassifier() - AmarendraGhanekarC5T3DTREEipynb
5. **Extra Trees** - ExtraTreesClassifier() - AmarendraGhanekarC5T3ETREE.ipynb

Based on the evaluation and comparison of all models, I selected Decision Tree model.

**Decision Tree Model Performance:**

**Binary confusion matrix D.TREE:**

Predicted False True \_\_all\_\_

Actual

False 577 167 744

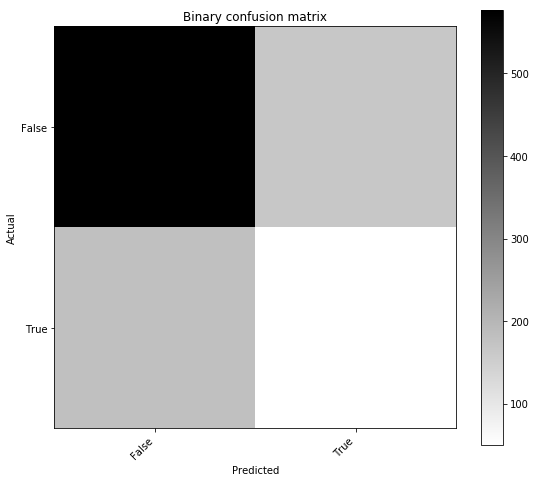
True 180 50 230

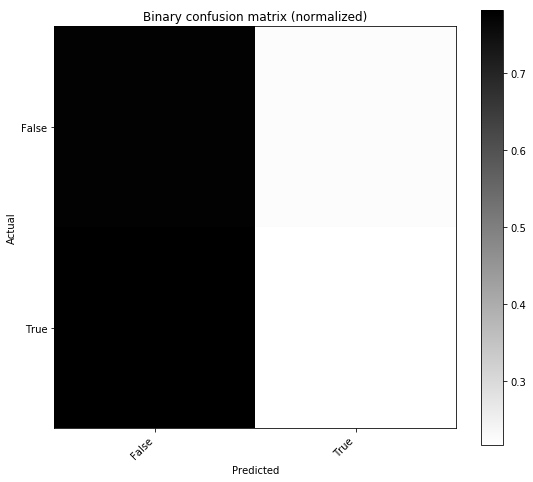
\_\_all\_\_ 757 217 974

**Accuracy Score** 0.7236116219198234

**Classification Report**

|  |
| --- |
| **precision recall f1-score support** |
|  |
| 0 0.82 0.82 0.82 4208 |
| 1 0.39 0.38 0.38 1230 |
|  |
| **avg / total 0.72 0.72 0.72 5438** |





# Observations and Recommendations

1. Although we cannot control customer spending habits, we can certainly provide features in the online billing system, which will allow customer to setup up their custom alerts for spending thresholds. This can be similar to the threshold system used by cable/wireless providers for data utilization. Customer awareness is the key to success.
2. Every customer has unique background, situation and habits. In addition, these variable change over time. Hence it is very difficult to predict exactly how a customer will behave in future, but the data can provide known patterns.
3. Further continuous analysis is needed to identify and monitor the features that impact the default in payments. The trends are likely to change depending upon socio/political, economic and regulatory changes. Personal attributes such as major changes in life, health, education can and will affect customer behavior. The root cause for default can be different for different people in different circumstances.
4. Although is difficult to accurately predict which customers can/will pay their loans, based on the identified patterns using selected model, we can proactively reach out to the candidates with options such as revised repayment plans.
5. The selected model will help us approve customers with higher certainty but continuous monitoring of data patterns are necessary because the variables do change over time as explained in #3 above.