|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Predicting Credit Risk  German Credit Data |
| |  |  |  | | --- | --- | --- | | Team#7: Shrinet, Nupur & Shetty, Aashna | 12/9/20 | Predictive Analytics (MSC ) | |

Explain what investigative questions you will be addressing and how these relate to possible policy recommendations.

# Inspiration & Introduction:

U.S. banks will set aside up to $320 billion to cover potential credit losses in 2020 due to the financial strain of the pandemic, according to a new report from Accenture (NYSE: ACN).

Chart, histogram

Description automatically generated

In 2019, U.S. banks set aside US$55 billion to cover potential loan losses; Accenture estimates that banks will need to hold an additional US$210 billion to US$265 billion to cover potential write-offs for bad loans in 2020 depending on the severity of the public health aspect of COVID-19. In Europe, banks might write off up to US$460 billion in 2020, an increase of US$370 billion from 2019; and in China, banks might write off up to US$360 billion in 2020, a US$190 billion increase from 2019.

And so now more than ever it is important for banks to evaluate what are the factors that result in a good credit. What type of loans should they encourage more and what type should they stop advertising etc.

The current report illustrates the data analysis of the behavior of the German borrowers. Multiple question such as *What is the duration of account with the bank? How many dependents does the individual have? What is the credit amount owned by the borrower?* The project answers many such question to understand the pattern for credit risk being good or bad. We will explore various features with respect to our target variable i.e. Credit Risk (V21) in our exploratory data analysis.

Investigative Problems we will address are:

* Predict Credit Risk based on behavior and characteristics of existing buyers.
* What are the factors that result in Bad Credit?
* What questions are mandatory to be answered while evaluating credit decisions?
* What variables should be advertised? For ex: Are customers who have current a/c relationship with the bank have lower credit risk? etc.
* What is the minimum threshold of credit score that the bank should accept while giving credit?
* Are mortgage-based credits lowering credit risks or vice versa?
* Are customers with multiple credits high risk?

Explain why certain methods for analysis have been selected. You should use at least two methods to develop results. Explain the model selection rationale and the comparative advantage & potential weakness for each of the chosen modeling techniques

# Model Selection:

Below were the steps that were used to decide the model.

1. As per the goal of the project, we found that we have a target variable in the data. This narrowed our selection between supervised and unsupervised algorithm to supervised learning.
2. We now have two type of supervised learning in front of us to select, i.e. Regression and Classification
3. After looking into the type of the target variable (i.e. binary/categorical) we will proceed with *Classification model/algorithm*, rejecting the regression model due to target variable being categorical
4. As we proceed further, we had plethora of classification model to choose. We decided to select one basic model and one advance/blended model to understand the lift that the blended model have against the basic model

Below were the two model that were selected against each model criteria (i.e. base and blended model)

1. **Logistic regression** (analysis performed using ***R***): We selected the logistic model for our base model compared to other decision tree models (C5.0, CHAID, Quest, CNR) because we are dealing with high dimensional data. Also learning of simple hyperplane in logistic or logit model is better due to the curse of dimensionality
2. **XG Boost Tree** (analysis performed using ***SPSS***): We selected XG Boost against the Random Forest model, as we had imbalanced target variable 70% positive and 30% negative. The XGBoost model comes from gradient boosting family of algorithm which is great for most of the classification models (specifically for imbalanced group). Also as the data only have 1000 record we will not have issues with slowness of building model using XGBoost

Below is the comparison of L*ogistic* and *XG boost* based on advantages and disadvantages

**XGBoost model**

Advantages of using XGBoost model

* XG Boost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from overfitting. That is why, XG Boost is also called regularized form of GBM (Gradient Boosting Machine).
* Parallel Processing
* Flexibility in setting an objective function
* Handling of missing values
* Non-Greedy tree pruning
* Built in Cross-Validation & methods of dropout (DART)

Potential weakness of using XGBoost

* XGBoost, have a requirement to manually create label encoding/ one hot encoding variable for categorical features as preprocessing step before feeding the features into the model. There are another gradient based model such as *lightgbm* that can do this by themselves
* Training time for XGBoost model on larger dataset is larger, if you compare against another gradient based model (e.g. *lightgbm*)
* As XGBoost involves boosting there is a chance that the model is overfit because of boosting

**Logistic regression model**

Advantages of using Linear Regression model

* Logistic regression is easy to interpret, implement and very efficient in training
* There is no assumption on distribution of levels/classes in feature space
* It not only provides the strength of the predictor (coefficients) but also the direction of association (positive or negative)
* It is less inclined toward overfitting

Potential weakness of using Linear regression

* In scenarios where number of features exceeds the number of observation logistic regression cannot be used
* Nonlinear type of business problem cannot be solved using logistic regression
* The model suffers from biasness if there is multicollinearity among the dependent variables

Specify the dataset source and its description. Develop a list of variables used in your modeling. Briefly describe any new data items derived from client-supplied data and any external data that you added to the source data provided. What if any limitations existed from the source data and potential implications for the quality of the results?

# Dataset Source:

**Context**

The dataset contains 1000 entries having 20 attributes prepared by Professor Dr. Hans Hofmann. The data classify individual by a set of attributes as good or bad credit risk (target variable). The attributes have both integer and categorical variables defining characteristics of an individual as good or bad credit.

**Content**

The original dataset has column headers coded as V1, V2……V21, and the value of each categorical column labeled as coded categorical values. Below are the descriptions of each categorical value codes.

***\*\*\*****Refer the Appendix section for the details of the variables and description*

**Limitation of Data**

The target value distribution is not equally distributed i.e. 700 (1- Good risk) and 300 (2-Bad risk).

This is a common issue in scenarios where anomaly detection is important such as fraudulent of transaction in banks, identification of disease in patient, etc. If we ignore the current issue, the predictive model created will be biased and inaccurate. The reason of this happening is because the mathematical models are designed in order to improve the accuracy by reducing the misclassification (error) and thus they do not consider the class disproportion or class balance.

Clean/transform/merge/select the data for analysis. What was done--why?

# Data Preparation (Clean/Transform/Merge/Select)

1. Imported data from var node and connected the type node. Converted field to Record Id and Credit Risk to Target. Further changed data types based on needs.
2. Used Reclassify Node to change original values of the variables to new values that are more understandable like a data dictionary. Further renamed all Predictors for ex like V1 which meant Current Ac Status to more clear names.
3. Attached data audit node to see data quality. There are no missing/Null/Blank/Whitespace data. We just have some Outliers. Coerced Outliers and generated an Outlier/extreme Super node.
4. Attached a Distinct node to check unique values by record id. And then further another Unique node to check on other variables. Attached tables to understand if the records are reduced, but our data has no duplicates.
5. We further use the filter function to filter out all continuous variables and use the auto data prep node on them to further make them ready for modelling. We set the auto data prep field to optimize accuracy. Our Auto data prep transforms the continuous variables as we see in the table. Transformation replaces the variable by a function of that variable: for example, replacing a variable x by the square root of x or the logarithm of x. In a stronger sense, a transformation is a replacement that changes the shape of a distribution or relationship.
6. We further use filter on our Ordinal variables and reclassify them so that we can label encode them. For ex: a person having a 5-year relationship of Saving acc with the bank is superior to someone whose saving acc is just one year old or does not have any relationship with the bank.
7. We then use the merge option to combine our categorical and continuous variables ready to be modelled.
8. We then use the Filler node to convert our categorical data that we had label encoded to be stored as a number instead of text so we could add that to the model. If we do not convert the storage of our label encoded ordinal variables, the model will give error as it must be stored as number instead of string. We further filter our nominal values to one hot encode them so that we can use them for modelling. We then merge the data we filtered out to one hot encode.
9. Graphing the distribution of our target variable. We see that there is imbalance with 70 % being good and 30% bad. We balance our data by a factor of 0.5 to reduce the sample. We also tested with boosting and modelling without balancing since XG boost takes care of unbalanced dataset but this proportion of under sampling with a factor of 0.5 for good credit gave the maximum accuracy considering the bias variance tradeoff. We then graph the credit risk again to check if our balancing worked.
10. We then partition the data into testing and training at 75% and 25% and import XG boost tree to prepare the model. We specify our target and predictors and run the model. We further Divide the result into a training and testing matrix.

Perform the DM analyses including models and additional evaluation of model output, e.g., web, gains, matrix as appropriate. Explain your results and any sensitivity analyses performed. At least 2 different models should be used in your project

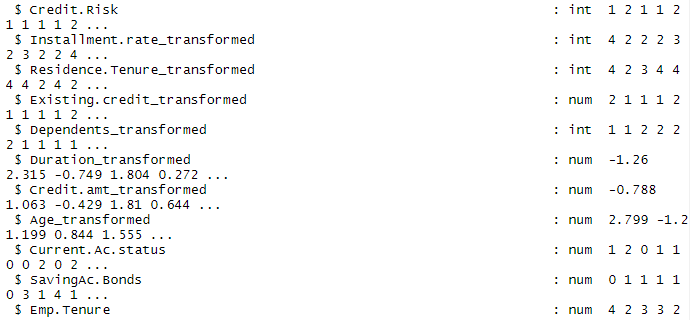
# Data Analysis and Model Configuration & Evaluation

**Data Analysis (refer R Word file for code)**

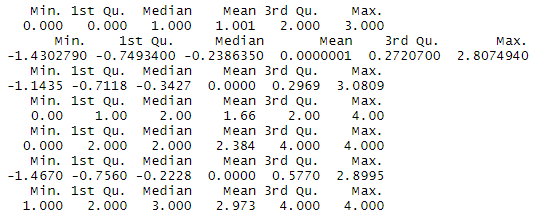
We wanted to explore the structure of the data followed by summary statistics and graphical representations of the variables to understand the dataset better. We understood the following from our initial analysis.

Structure and Summary statistics indicated the presence of categorical variables as well as nominal variables and showed no outliers. The continuous variables that were considered were: Duration, Current Ac Status, Credit History, Credit Amt, Emp Tenure, Age, and Installment rate. Nominal variables were Property owned, Purpose, Status and Sex (Note that dummy variables were created for the levels of each nominal variable).

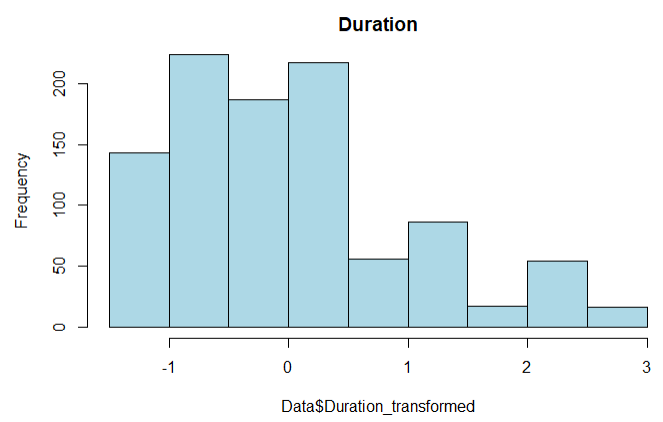
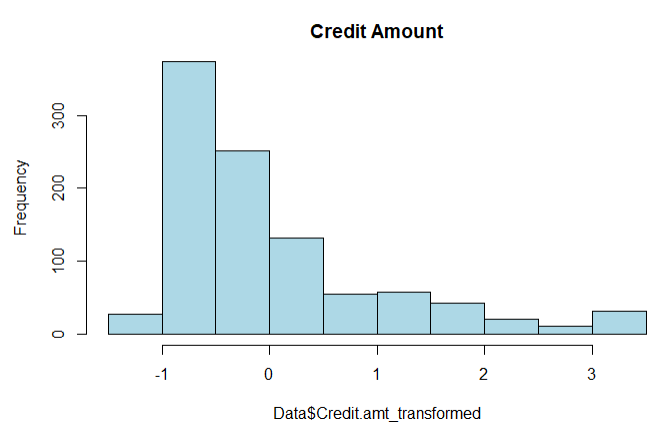
Structure of Data

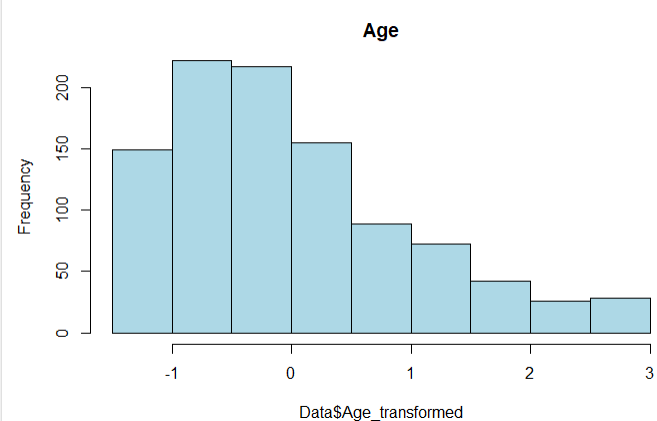
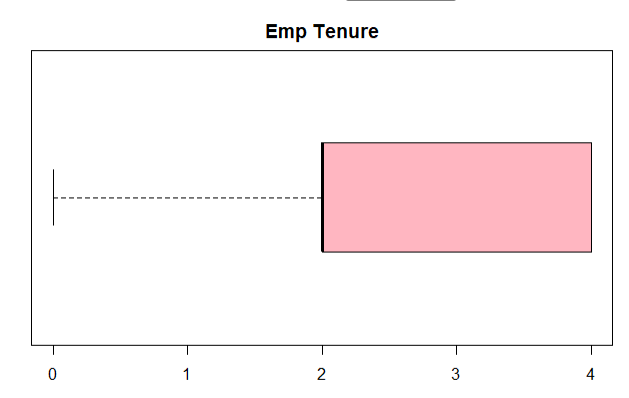


Summary Statistics

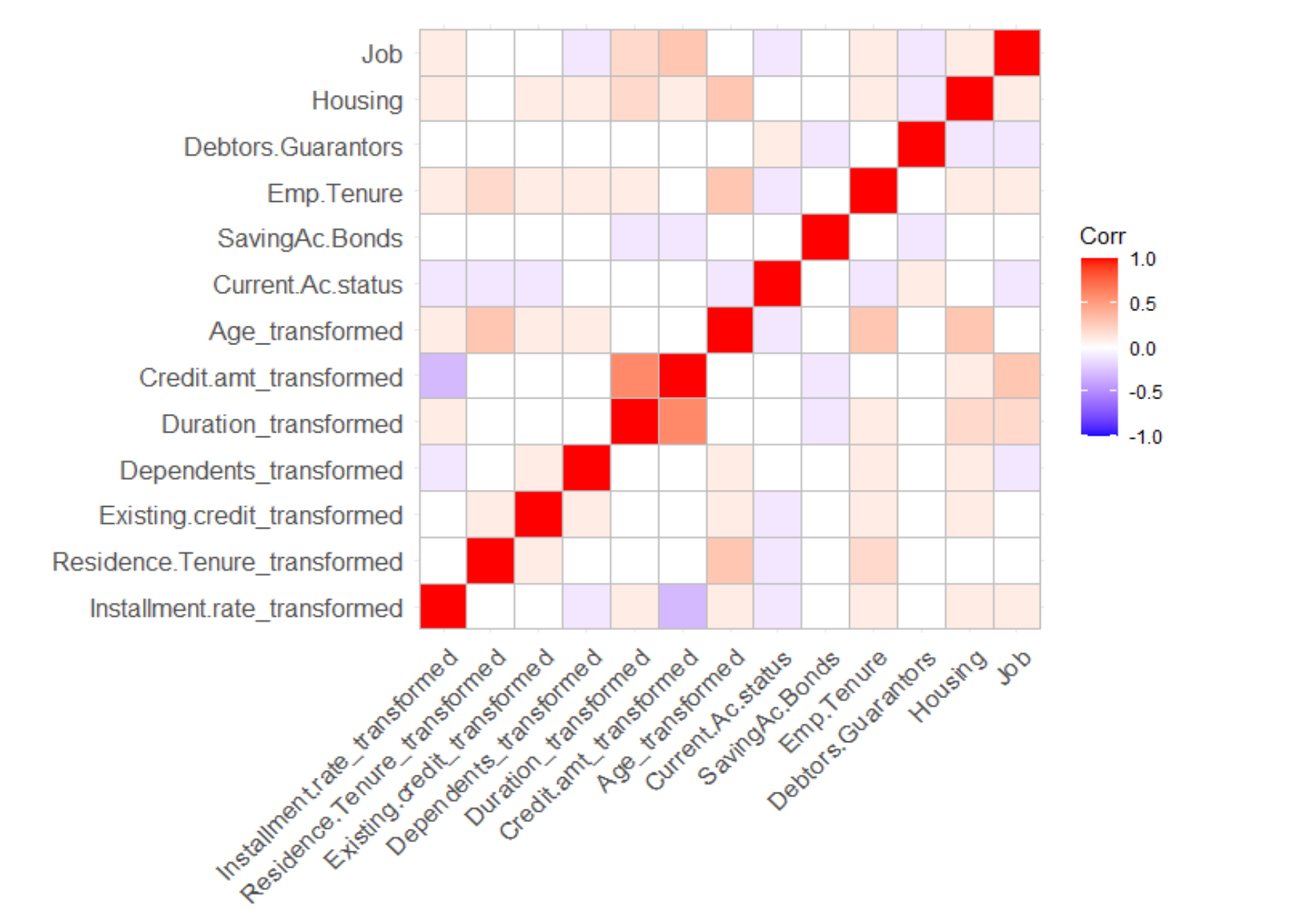


We plotted Histograms and Box plots to understand the distribution of the data visually. Many variables were considered but here are a few interesting ones. (Note the knitted R file will have graphs for each variable considered, below are just a few examples)





Lastly, we wanted to check how each of these variables interacted with each other. Therefore, we plotted a scatter plot matrix.



**Configuration & Evaluation of Models:**

We tried balancing the data using the Balance node and then connected the balanced node to the models. The focus of the predictive model will be to positive class (bad credit). Good starting point will be precision and recall, we cannot rely on accuracy because the **cost** of predicted a bad credit wrong is more than other class . If we maximize precision the false positive will be reduced and similarly maximizing the recall will also reduce the false negative in the prediction model.

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

We can ultimately also use AUC curve to define the accuracy and performance of model

* **XGBoost Model (SPSS Modeler, refer the SPSS stream):** The XG Boost model was ran using all predefined roles of the features and below metric were created on Training and Testing dataset using the Analysis tool.

Comparing $XGT-Credit Risk with Credit Risk

'Partition' 1\_Training 2\_Testing

Correct 405 84.55% 139 83.73%

Wrong 74 15.45% 27 16.27%

**Total 479 166**

*Note\* The above matrix has less observation because of balancing of data.*

Coincidence Matrix for $XGT-Credit Risk (rows show actuals)

'Partition' = 1\_Training 1 2

1 219 37

2 37 186

'Partition' = 2\_Testing 1 2

1 71 18

2 9 68

We will use above coincidence matrix of both Training and Testing to calculate the Precision and Recall of the model

**Training**

Precision = 186/ (186+37) = 83.4%

Recall (Sensitivity) = 186/ (186+37) = 83.4%

F-Score = 2\*Recall\*Precision/ (Recall + Precision) = 0.834

AUC = 0.891

**Testing**

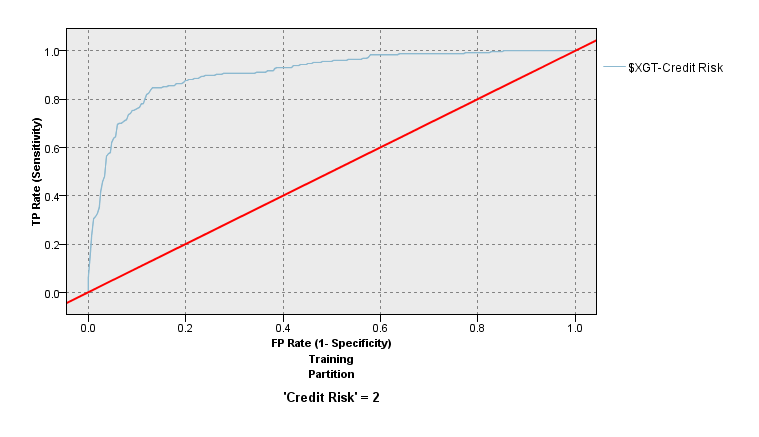
Precision = 68/ (68+18) = 80%

Recall (Sensitivity) = 68/ (68+9) = 88.3%

F-Score = 2\*Recall\*Precision/ (Recall + Precision) = 0.838

AUC = 0.90

The model has a good AUC value and perform equal on both testing and Training dataset. This explains that the model is consistent and stable with great performance.



ROC curve on training data

* **Logistic Regression (R, refer the R Notebook file)** The Logistic regression model was ran using all the features from cleaned data and below metric were created on Training and Testing dataset using the R code.

Comparing $XGT-Credit Risk with Credit Risk

'Partition' 1\_Training 2\_Testing

Correct 329 70.6% 144 68.2%

Wrong 137 29.4% 67 31.8%

**Total 466 211**

*Note\* The above matrix has less observation because of under balancing of data.*

Coincidence Matrix for $XGT-Credit Risk (rows show actuals)

'Partition' = 1\_Training 1 2

1 165 68

2 69 164

'Partition' = 2\_Testing 1 2

1 94 50

2 17 50

We will use above coincidence matrix of both Training and Testing to calculate the Precision and Recall of the logistic model

**Training**

Precision = 164/ (164+68) = 70.6%

Recall (Sensitivity) = 164/ (164+69) = 70.4%

F-Score = 2\*Recall\*Precision/ (Recall + Precision) = 0.705

**Testing**

Precision = 50/ (50+50) = 50%

Recall (Sensitivity) = 50/ (50+17) = 74.6%

F-Score = 2\*Recall\*Precision/ (Recall + Precision) = 0.598

The model has a good performance equal on both testing and training dataset. This explains that the model is consistent and stable with great performance.

Discuss what you learned to confirm or disconfirm expectations any meaningful patterns discovered from predictive analytics and/or EDA methods (e.g., cluster analysis). What 3 or more recommendations are proposed to your client?

# Finding/Recommendation

Findings

* Feature: Duration in month – Increase in duration of month increase the odd of being a person default to loan
* Feature: Installment plan – Approving a loan without any installment plan reduces the default on loans
* Feature: Other Debtors – Approving the loan application with guarantor decrease the odd of defaulting on the loan
* Feature: Dependents – People with more dependents in family are more likely to default on loan than others
* Feature: Credit History - People who have taken credit and completed the payment have less risk compared to other profiles
* Model: XG boost is the best classifier among the two models with better precision and recall(sensitivity)
* Model: The models were stable and robust based on the performance on training and testing datasets

Recommendations

* Cluster the applicant into the three categories of risk i.e. low, medium and high risk the applicant should be categorized based on loan amount and other demographics characteristics
* Average income per family and loan amount requested should be most important factor in decision of approving the loan
* High loan amount request can be used a filtering engine as first pass on application, the filtered application can then go through the rigorous process to select applicants
* The project can be extended to assign different weights to each of the features as part of feature engineering process before building and training the classifier
* Use of external data sources: Modern Credit/Fraudulent analysis involves using other data sources such as criminal record of applicant, health insurance and health information, outstanding balance out of monthly income and expenses

# Appendix

**Target(categorical):**

*Label*: V21 *Descriptive Column Name*: Credit risk

Value Mapping

1: Good Credit  
2: Bad Risk

**Feature 1(categorical):**

*Label*: V1 *Descriptive Column Name*: Status of existing checking account

Value Mapping

A11: ... < 0 DM  
A12: 0 <= ... < 200 DM  
A13: ... >= 200 DM / salary assignments for at least 1 year  
A14: no checking account

**Feature 2(numerical):**

*Label*: V2 *Descriptive Column Name*: Duration in months

**Feature 3(categorical):**

*Label*: V3 *Descriptive Column Name*: Credit history

Value Mapping

A30 : no credits taken/ all credits paid back duly  
A31 : all credits at this bank paid back duly  
A32 : existing credits paid back duly till now  
A33 : delay in paying off in the past  
A34 : critical account/ other credits existing (not at this bank)

**Feature 4(categorical):**

*Label*: V4 *Descriptive Column Name*: Purpose of credit

Value Mapping

|  |  |  |
| --- | --- | --- |
| A40 : car (new) | A41 : car (used) | A42 : furniture/equipment |
| A43 : radio/television | A44 : domestic appliances | A45 : repairs |
| A46 : education | A47 : (vacation - does not exist?) | A48 : retraining |
| A49 : business | A410 : others |  |
|  |  |  |

**Feature 5(numerical):**

*Label*: V5 *Descriptive Column Name*: Credit amount

**Feature 6(categorical):**

*Label*: V6 *Descriptive Column Name*: Saving accounts/bonds

Value Mapping

A30 : no credits taken/ all credits paid back duly  
A31 : all credits at this bank paid back duly  
A32 : existing credits paid back duly till now  
A33 : delay in paying off in the past  
A34 : critical account/ other credits existing (not at this bank)

**Feature 7(categorical):**

*Label*: V7 *Descriptive Column Name*: Employment tenure

Value Mapping

A71 : unemployed  
A72 : ... < 1 year  
A73 : 1 <= ... < 4 years  
A74 : 4 <= ... < 7 years  
A75 : .. >= 7 years

**Feature 8(numerical):**

*Label*: V8 *Descriptive Column Name*: Installment rate in % of disposable income

**Feature 9(categorical):**

*Label*: V9 *Descriptive Column Name*: Personal status and sex

Value Mapping

A91 : male : divorced/separated  
A92 : female : divorced/separated/married  
A93 : male : single  
A94 : male : married/widowed  
A95 : female : single

**Feature 10(categorical):**

*Label*: V10 *Descriptive Column Name*: Other debtors / guarantors

Value Mapping

A101 : none  
A102 : co-applicant  
A103 : guarantor

**Feature 11(numerical):**

*Label*: V11 *Descriptive Column Name*: Present residence since

**Feature 12(categorical):**

*Label*: V12 *Descriptive Column Name*: Property Owned

Value Mapping

A121 : real estate  
A122 : if not A121 : building society savings agreement/ life insurance  
A123 : if not A121/A122 : car or other, not in attribute 6  
A124 : unknown / no property

**Feature 13(numerical):**

*Label*: V13 *Descriptive Column Name*: Age in years

**Feature 14(categorical):**

*Label*: V14 *Descriptive Column Name*: Other installment plans

Value Mapping

A141 : bank  
A142 : stores  
A143 : none

**Feature 15(categorical):**

*Label*: V15 *Descriptive Column Name*: Housing

Value Mapping

A151 : rent  
A152 : own  
A153 : for free

**Feature 16(numerical):**

*Label*: V16 *Descriptive Column Name*: Number of existing credits at this bank

**Feature 17(categorical):**

*Label*: V17 *Descriptive Column Name*: Job

Value Mapping

A171 : unemployed/ unskilled - non-resident  
A172 : unskilled - resident  
A173 : skilled employee / official  
A174 : management/ self-employed/  
highly qualified employee/ office

**Feature 18(numerical):**

*Label*: V18 *Descriptive Column Name*: Number of people being liable to provide maintenance for

**Feature 19(categorical):**

*Label*: V19 *Descriptive Column Name*: Telephone

Value Mapping

A191 : none  
A192 : yes, registered under the customers name

**Feature 20(categorical):**

*Label*: V20 *Descriptive Column Name*: foreign worker

Value Mapping

A201 : yes  
A202 : no