Credit Card Risk Classification

Members

Yun Shiuan Hwang, yun.hwang@ryerson.ca
Ann Sam-Erb, ann.sam@ryerson.ca
Khilan Shah, khilan.shah@ryerson.ca

Contents

Summary

Good credit or bad credit? That is the question. Using the German Credit Dataset, our team has developed a data analytics-based strategy to help classify if prospective borrowers would be a risk to the client bank. As many applications are submitted and processed every day, predictive modeling would help assist the bank executives in determining whether an applicant is approved or rejected for a loan.

The German Credit dataset consists of 1000 records (rows) with 21 attributes (columns). One class attribute of 'Creditability' is our target variable of Good Credit = 1 and Bad Credit = 0. The 20 other attributes are other predictors such as Duration of Credit, Payment Status of Previous Credit, Purpose, Credit Amount, etc. Based on the information within the dataset, our team utilized Python 3.7 and third-party packages such as scikit-learn for our predictive modeling. A Decision Tree model was generated with results showing an accuracy of 70% and the use of a Naive Bayes classifier with results showing an accuracy of 63%.

In the data preparation process, we identified 4 attributes with low correlation for the predictive analysis and during the post-predictive analysis, a filtered dataset was created with only the 17 of the more correlated predictors. After rerunning the filtered dataset, the decision tree classifier returned an increase in accuracy from 70% to 74%. For the Naive Bayes classifier, it also returned an increase in accuracy from 63% to 64%.

The results of the predictive analysis leads us to recommend that the client bank use the 17-attribute filtered data set as predictors for future borrowers. The increase in accuracy supports the idea that while processing lesser predictors, their stronger correlation is more beneficial to ensuring accuracy rather than processing weaker correlated data that could introduce errors or dilution in applicant data.

Workload Distribution

Member Name	List of Tasks Performed
Yun Shiuan Hwang	Post Predictive Analysis, Conclusion

Ann Sam-Erb	Summary, Data Prep, Post Predictive Analysis
Khilan Shah	Predictive Modeling, Post Predictive Analysis

Data Preparation

Dataset Summary:

The German Credit Dataset was provided in two formats: comma-separated values (CSV) and Attribute-Relation File Format (ARFF). The dataset contained 1000 records. All categorical values were already converted to indicator values and a search for null or missing values returned that the dataset was complete. The target/class variable is Creditability which shows whether the applicant has a good or bad credit rating. Within the dataset, 753 records are “Good Credit” while 247 records are “Bad Credit". The data was imported into Python 3.7 and a summary of the attributes within the dataset is provided below in Figure 1.

	Attributes	**Type**	**Min**	**Max**	**Mean**	**Standard Deviation**	**Distinct Values**	**Missing Values**
1	Creditability	nominal	-	-	-	-	2	0
2	Account Balance	ordinal	-	-	-	-	4	0
3	Duration of Credit (month)	quantitative	4.00	72.00	20.90	12.06	33	0
4	Payment Status of Previous Credit	ordinal	-	-	-	-	5	0
5	Purpose	nominal	-	-	-	-	10	0
6	Credit Amount	quantitative	250.00	18424.00	3271.25	2822.75	923	0
7	Value Savings/Stocks	ordinal	-	-	-	-	5	0
8	Length of current employment	ordinal	-	-	-	-	5	0

9	Installment per cent	quantitative	1.00	4.00	2.97	1.12	4	0
10	Sex & Marital Status	nominal	-	-	-	-	4	0
11	Guarantors	nominal	-	-	-	-	3	0
12	Duration in Current address	ordinal	-	-	-	-	4	0
13	Most valuable available asset	nominal	-	-	-	-	4	0
14	Age (years)	quantitative	19.00	75.00	35.54	11.35	53	0
15	Concurrent Credits	nominal	-	-	-	-	3	0
16	Type of apartment	nominal	-	-	-	-	3	0
17	No of Credits at this Bank	quantitative	1.00	4.00	1.41	0.58	4	0
18	Occupation	nominal	-	-	-	-	4	0
19	No of dependents	quantitative	1.00	2.00	1.16	0.36	2	0
20	Telephone	nominal	-	-	-	-	2	0
21	Foreign Worker	nominal	-	-	-	-	2	0

Table 1. Attribute summary of the German Credit dataset.

Hot-one coding was not required during the data preparation process as the categorical data was already converted into numerical values to represent the categories. The list of attributes and their categorical conversions are as follows:

Creditability: 1: Good Credit 2: Bad Credit	**Credit Amount: Numerical value showing the credit amount**	**Most valuable available asset: Qualitative attribute showing valuable assets** 1: real estate 2: savings agreement/ life insurance 3: car or other 4: unknown / no property
Account Balance: 1: < 0 DM, 2: 0<=...<200 DM 3: > 200 DM 4: No checking accounts *where DM= Deutsche Mark*	**Value Savings/Stocks:** 1: <100 DM 2: 100<= ... < 500 DM 3: 500<= ... < 1000 DM 4: =>1000 DM 5: unknown/ no savings account	**Age (years): Numerical value showing age in years.** **Concurrent Credits: Installment plans** 1: bank 2: stores 3: none
Duration of Credit (month)	**Length of current employment:** 1: unemployed 2: < 1 year 3: 1<=...<4 years 4: 4<=...<7 years 5:>=7years	**Type of apartment: Type of housing** 1: rent 2: own 3: for free
Duration of credit in months (numerical)	**Installment percent: Installment rate in percentage of disposable income (numerical)**	**No of Credits at this Bank: Numerical value showing number of existing credits at the bank**
Payment Status of Previous Credit: 0: no credits taken 1: all credits at this bank paid back duly 2: existing credits paid back duly till now 3: delay in paying off in the past 4: critical account.	**Sex & Marital Status:** 1: male: divorced/separated 2: female: divorced/separated/married 3: male: single 4: male: married/widowed 5: female: single	**Occupation:** 1: unemployed/ unskilled - non-resident 2: unskilled - resident 3: skilled employee / official 4: management/ self-employed/highly qualified employee/ officer
Purpose: 0: New car 1: Used car 2: Furniture/Equipment 3: Radio/Television 4: Domestic Appliances 5: Repairs 6: Education 7: Vacation 8: Retraining 9: Business 10: Others	**Guarantors: Guarantors and co-applicants:** 1: none 2: co-applicant 3: guarantor	**No of dependents: Numerical value showing number of dependents**
	Duration in Current address: Qualitative value showing the duration in current address 1: <= 1 year 2: 1<...<=2 years 3: 2<...<=3 years 4: >3 years	**Telephone: Qualitative attribute for telephone number** 1: yes 2: No
		Foreign Worker: Qualitative attribute showing whether the person is a foreign worker 1: yes 2: no

Table 2. One-hot coding prepared in the given data set.

Visual Exploratory Data Analysis:

Looking at the 3 continuous-quantitative variables, Duration of Credit, Credit Amount, and Age (years), we first plot the data as a histogram to understand the distribution.

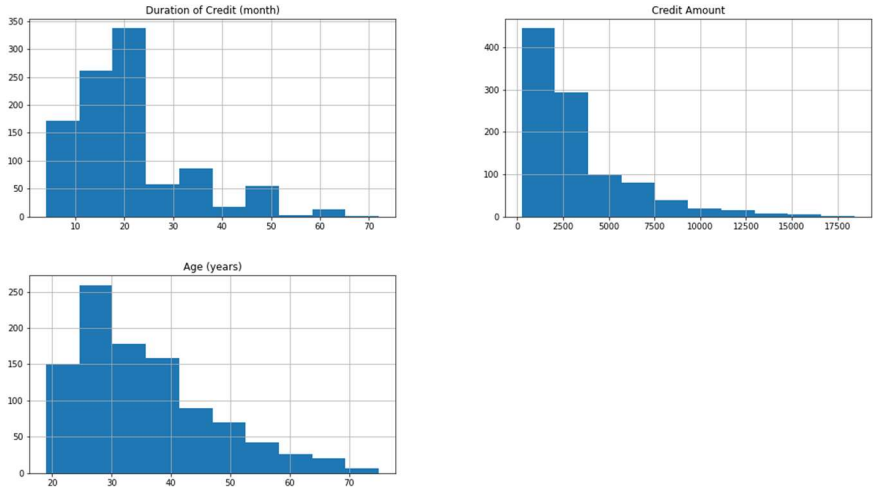


Figure 1. Histograms of continuous-quantitative variables.

The histogram of these attributes tells us that they all have slightly right-skewed distributions. While an ideal outcome is a bell curve, the skewed distribution is acceptable without having to reject any data.

We can plot these same variables against our target class variable, Creditability in a box-plot to view any outlier data.

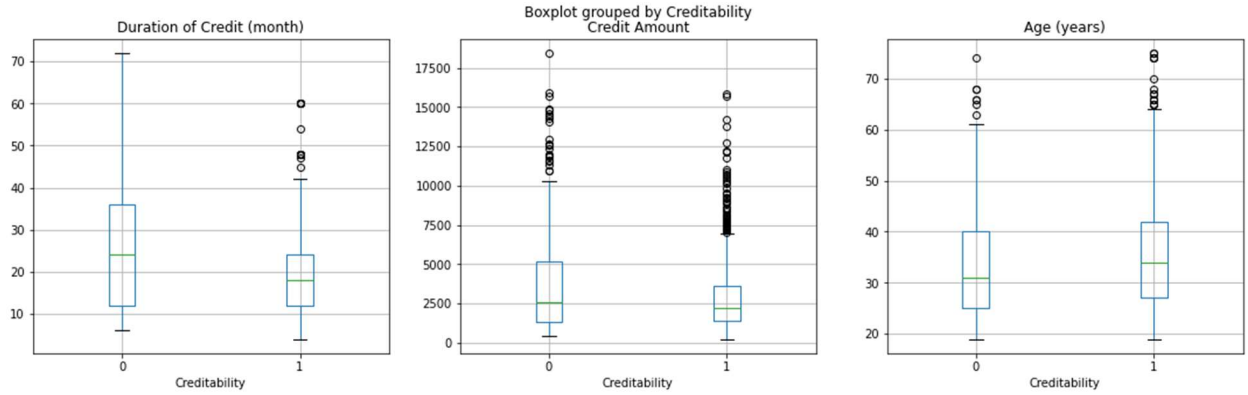
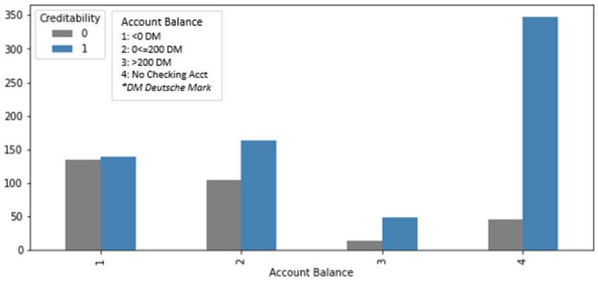
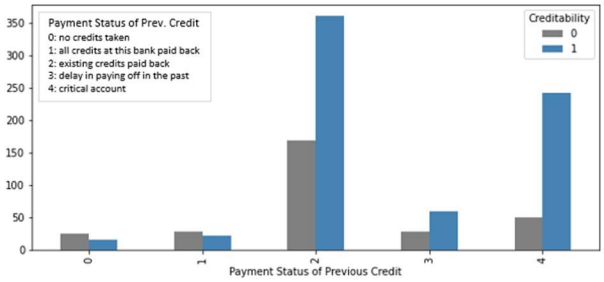


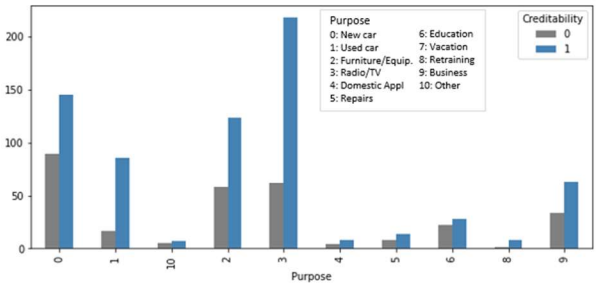
Figure 2. Boxplots of continuous-quantitative variables.

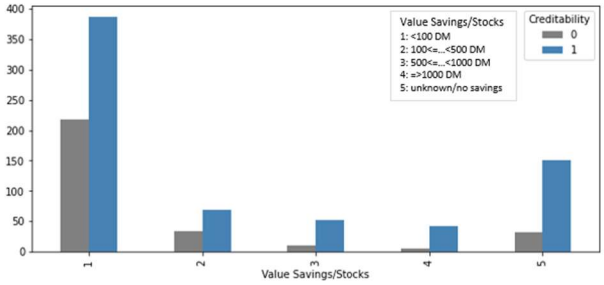
The boxplots show us the data distribution of the 3 continuous predictors in the y-axis of each variable. When interpreting the Duration of Credit and Credit Amount box plots, we can see that the boxes differ between the target variable which indicates that there is a correlation between those variables and the class variable: Creditability. Alternatively, we can see in the Age (years) boxplot, that the boxes are very similar in size and location within the plot and thus cannot explicitly differentiate between age as a good indicator of good or bad credit.

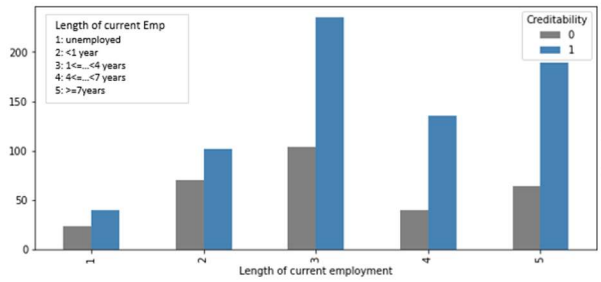
Using grouped Bar Graphs, we plotted the remaining 17 variables against our target class variable Creditability. The frequency in the y-axis and the categorical variable in the x-axis. To interpret the bar graphs, if we see that the ratio between the good and bad credit values are similar between columns, they may not be correlated to our target variable.

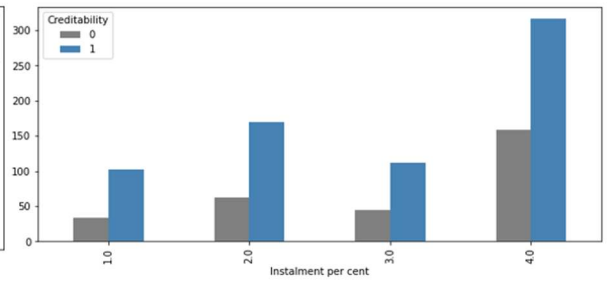

Creditability
0
1
Account Balance
1: <0 DM
2: 0<=200 DM
3: >200 DM
4: No Checking Acct
*DM Deutsche Mark
Account Balance

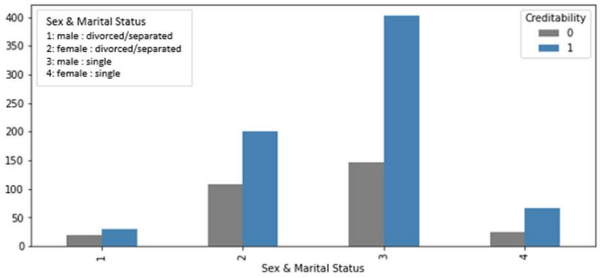

Payment Status of Prev. Credit
0: no credits taken
1: all credits at this bank paid back
2: existing credits paid back
3: delay in paying off in the past
4: critical account
Creditability
0
1
Payment Status of Previous Credit

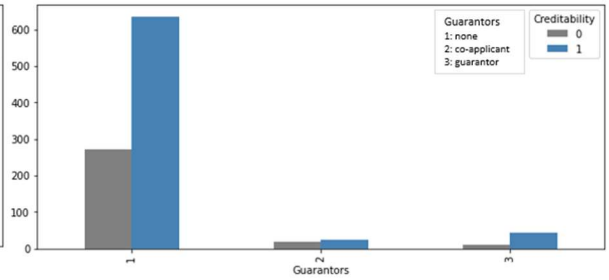

Purpose
0: New car
1: Used car
2: Furniture/Equip.
3: Radio/TV
4: Domestic Appl
5: Repairs
6: Education
7: Vacation
8: Retraining
9: Business
10: Other
Creditability
0
1
Purpose

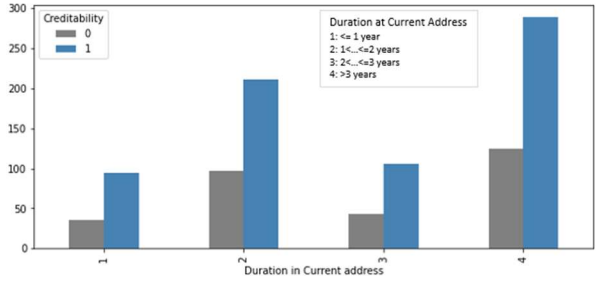

Value Savings/Stocks
1: <100 DM
2: 100<=...<500 DM
3: 500<=...<1000 DM
4: =>1000 DM
5: unknown/no savings
Creditability
0
1
Value Savings/Stocks


Length of current Emp
1: unemployed
2: <1 year
3: 1<=...<4 years
4: 4<=...<7 years
5: >=7years
Creditability
0
1
Length of current employment


Creditability
0
1
Instalment per cent


Sex & Marital Status
1: male : divorced/separated
2: female : divorced/separated
3: male : single
4: female : single
Creditability
0
1
Sex & Marital Status


Guarantors
1: none
2: co-applicant
3: guarantor
Creditability
0
1
Guarantors


Creditability
0
1
Duration at Current Address
1: <= 1 year
2: 1<...<=2 years
3: 2<...<=3 years
4: >3 years
Duration in Current address

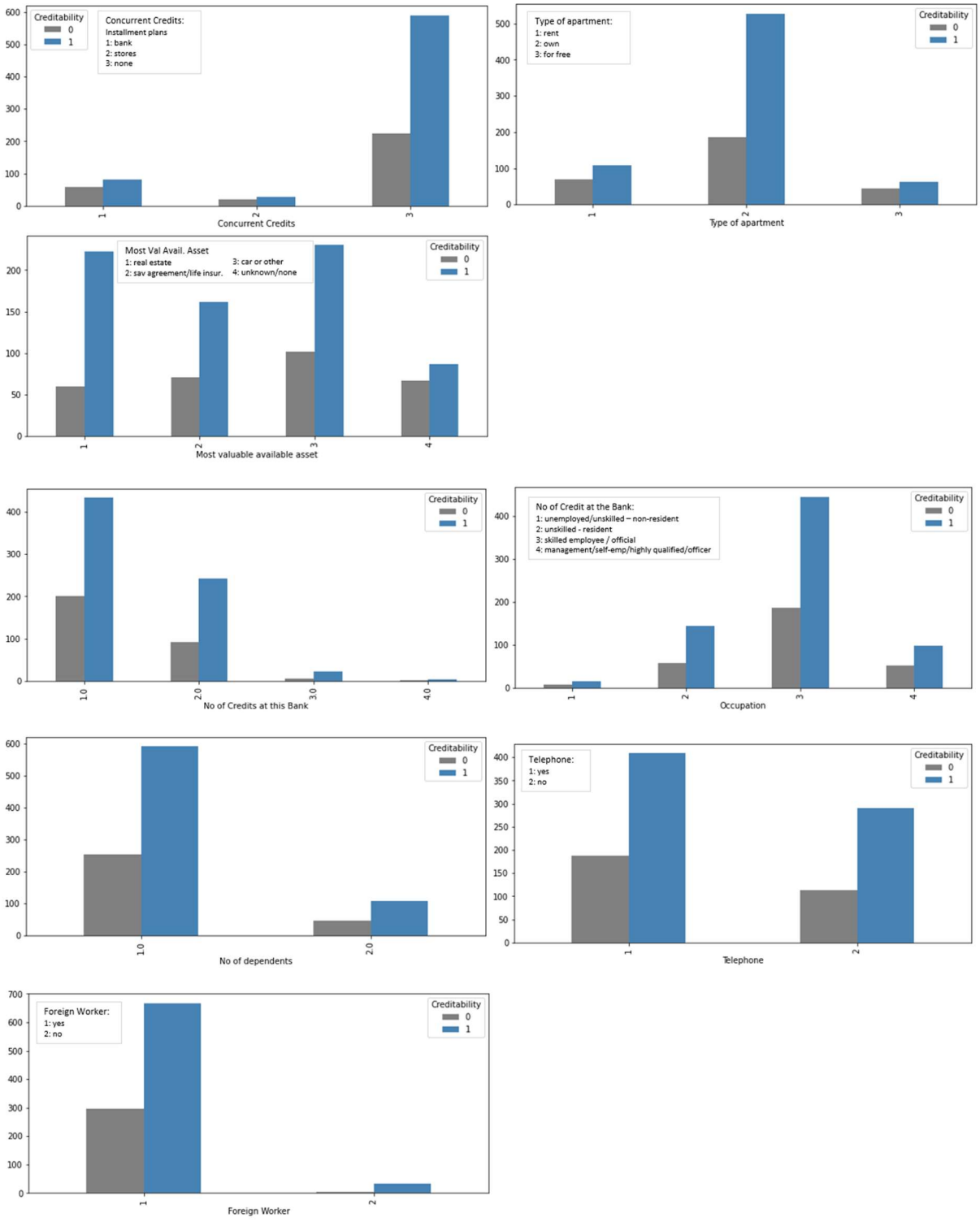


Figure 3. Bar graphs of categorical and discrete variables.

Based on analyzing the bar graphs, we can infer that there is a correlation between 'Creditability' and the following attributes: Account Balance, Payment Status of previous credit, Purpose, Value Savings/Stocks, Length of current employment, Sex & Marital Status, Guarantors, Most valuable available asset, Concurrent Credits, Type of apartment, and Foreign Worker. We should

note that when looking at the distribution of data, Foreign Worker data is skewed with very few records of applicants being foreign workers.

Statistical Attribute Analysis:

To confirm our visual analyses, we can apply statistical methods to the quantitative variables and the categorical variables.

We can apply a correlation function to the quantitative variables using the Analysis of Variance (ANOVA) test. The ANOVA test is performed by comparing two types of variation, the variation between the sample means, as well as the variation within each of the samples. The results of the ANOVA formula produce an F-statistic that allows for the analysis of multiple groups of data to determine the variability between samples. We can then use the p-value results from the calculation to determine if there is any correlation. A p-value less than 0.05 significance allows us to reject the null hypothesis which is: that there is no correlation.
(https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/)

	Continuous Attributes	Correlation with target	P-Value
1	Duration of Credit (month)	Yes	6.49E-12
2	Credit Amount	Yes	8.80E-07
3	Age (years)	Yes	3.87E-03

Table 3. ANOVA test correlation results between continuous variables and target variables.

After computing our 3 continuous variables through Python, we can statistically deduce that all three have a correlation with our target variable, Creditability.

For the categorical variables, we can apply Pearson's Chi-Squared method to compare the categorical variables with the target variable, Creditability. The chi-squared test is a statistical test which determines if the difference between 2 categorical variables is due to chance, or if it is due to a relationship between them. The assumption is that the columns are not related, similar to the ANOVA test previously mentioned. It then returns the calculated statistics and p-value for interpretation.
(https://www.analyticsvidhya.com/blog/2021/06/decoding-the-chi-square-test%E2%80%8A-%E2%80%8Ause-along-with-implementation-and-visualization/)

	Categorical Attributes	Correlation with target	P-Value
1	Account Balance	Yes	1.22E-26
2	Payment Status of Previous Credit	Yes	1.28E-12
3	Purpose	Yes	1.16E-04
4	Value Savings/Stocks	Yes	2.76E-07
5	Length of current employment	Yes	1.05E-03
6	Installment per cent	No	1.40E-01
7	Sex & Marital Status	Yes	2.22E-02
8	Guarantors	Yes	3.61E-02
9	Duration in Current address	No	8.62E-01
10	Most valuable available asset	Yes	2.86E-05
11	Concurrent Credits	Yes	1.63E-03
12	Type of apartment	Yes	8.81E-05
13	No of Credits at this Bank	No	4.45E-01
14	Occupation	No	5.97E-01
15	No of dependents	No	1.00E+00
16	Telephone	No	2.79E-01
17	Foreign Worker	Yes	1.58E-02

Table 4. Chi-Square test correlation results between continuous variables and target variables.

Decision on Attribute Filtering:

Based on visual and statistical exploratory analysis of the dataset, we can note that the following attributes appear to have little or no correlation with the target variable, Creditability:
1. Installment percent
2. Duration in Current address
3. No of Credits at this Bank
4. No of dependents

The original dataset will contain all 21 variables (inclusive of the target variable) and the filtered dataset will contain 17 variables (inclusive of the target variable).

Summary of Data Preparation:

The following steps were used to prepare the dataset for predictive modeling/classification:

1. The dataset was imported into Python 3.7 and assessed for null or na values.
2. The attributes/variables were classified into types for data processing
3. Using visual analytic processes, box-plots, histograms, and bar graphs were created.
4. To confirm our subjective analyses, statistical analyses were performed using the ANOVA and Chi-Square tests to determine correlations between the predictor variables and the target variable.
5. A filtered dataset was determined from step 4 processes.

As the dataset was complete without any missing or erroneous data. There was no need to complete any data normalization or transformation to prepare the data for modeling. The dataset did appear to have some skewed distributions but is acceptable to move forward with processing. After some statistical analysis on the correlation of the variables and the target variable, it was determined that there were 4 variables that had little or no correlation with determining good or bad credit. Those 4 variables will be omitted in a second filtered dataset that will also be used during the predictive modeling/classification process.

4. Predictive Modeling/Classification

Data Split Strategy

The train-test split is a technique for evaluating the performance of a machine learning algorithm. The strategy in our predictive analysis is to create a train-test split set, where 80% of the data set would be the training subset and the remaining 20% would be the testing subset. We will use the training data to train our models and use the test data to make sure the predictions are corroborated. The target class value to determine whether a customer has good credit the response will be ‘1’ and the response will be ‘0’ for bad credit.

Decision Tree

A decision tree is a flowchart-like structure where we can represent the attributes of the data set as a node and the branches represent a decision rule to another node and eventually to a leaf node that represents an outcome, creditability. The first step in our predictive modeling will be creating an algorithm and classifying the data using the Decision Tree Algorithm. In creating the decision tree, we will be splitting out data into training sets and testing sets as previously mentioned.

The depth parameter used for the decision tree was set to depth = 5 so we can get a better idea of how our data can be visualized. By setting the maximum depth to 5, it will allow the tree to set 5 splits starting with the root node and splitting the data by each attribute until a leaf node is achieved. The first attribute taken by the decision tree is Account Balance as it has the highest information gain, smallest entropy (impurity), and lowest GINI score. From that root node, it will be split into Duration of Credit (months), Lengths of Employment, etc.

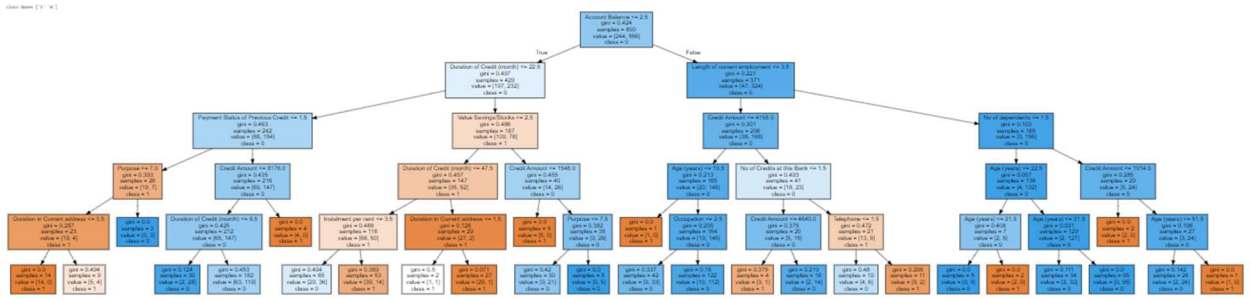


Figure 4. Decision Tree for original data set.

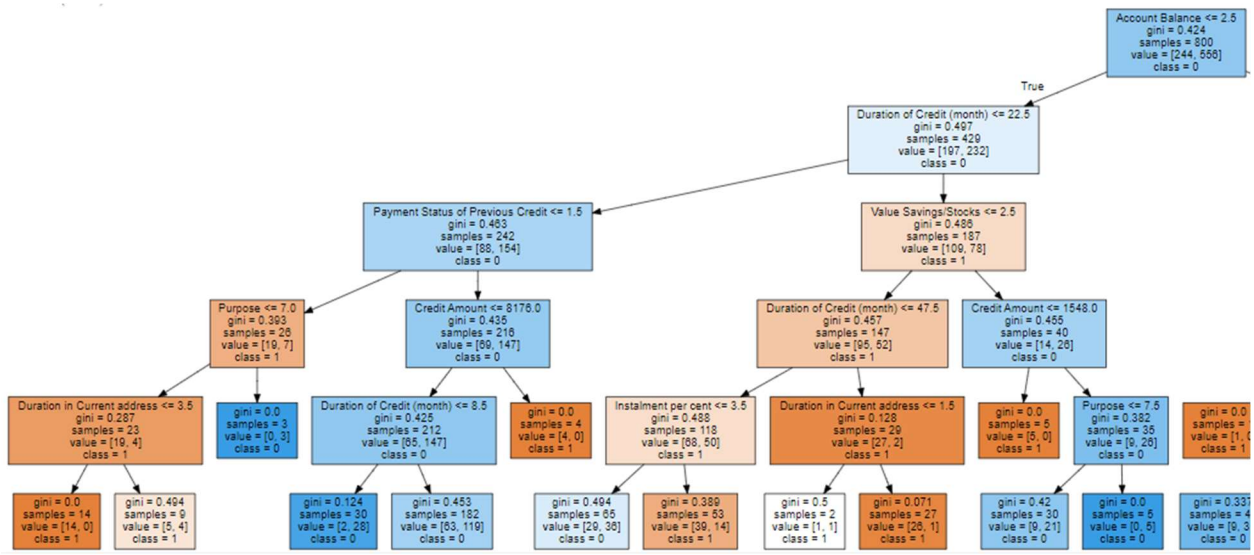


Figure 5. Zoomed in view of the left-side of theDecision Tree (original data set).

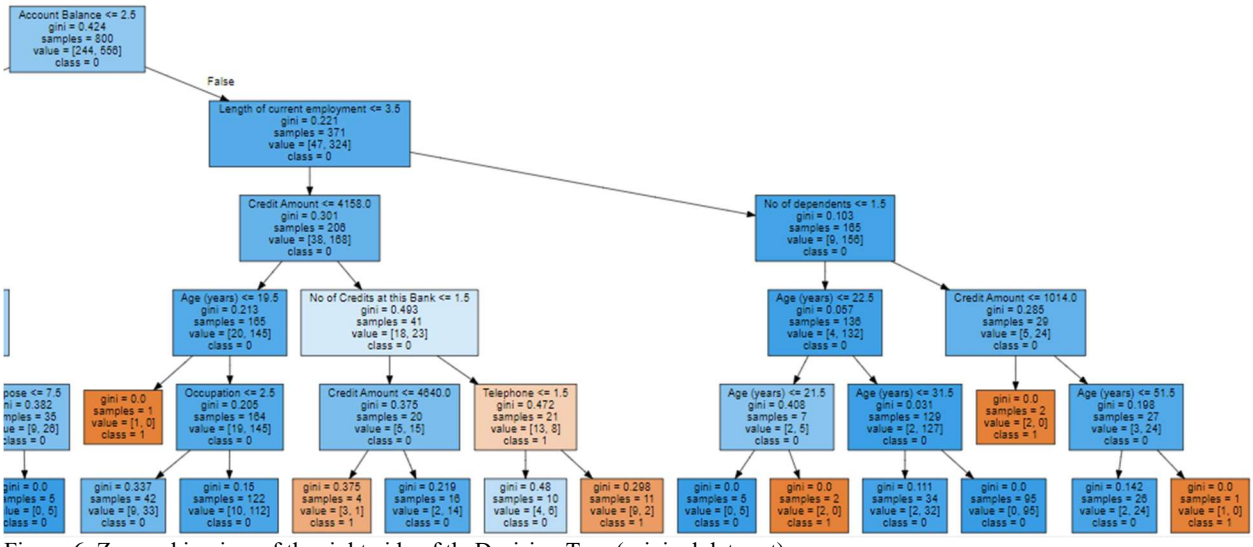


Figure 6. Zoomed in view of the right-side of theDecision Tree (original data set).

One interpretation of the decision tree can be summarized where a client has an Account Balance with 2.5 or less score, Duration of Credit (months) is less than 22.5, Payment Status of Previous Credit is less than 1.5, and the Credit Amount is less than 8176.0. An applicant with records that follow this path would be granted a credit card (i.e. have good credit). (Figure 5)

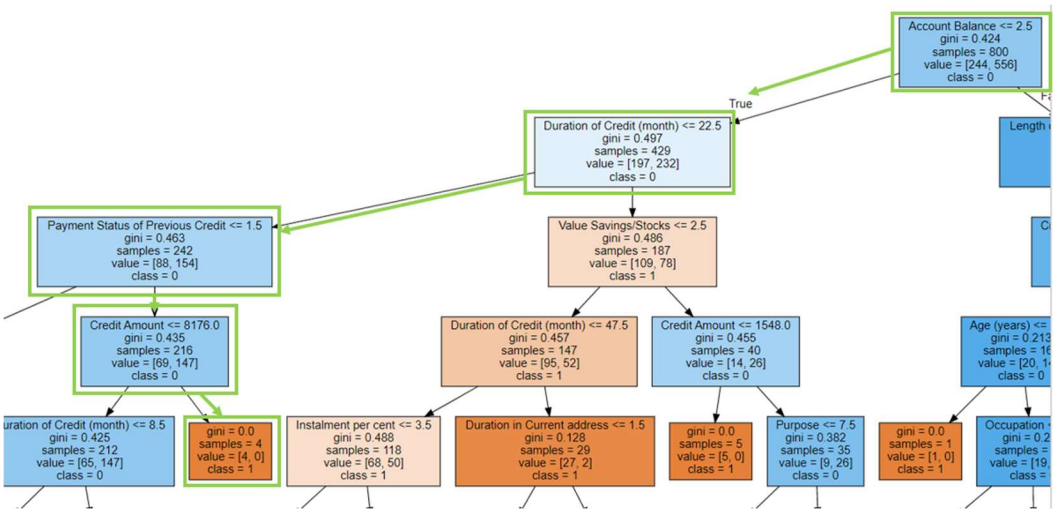


Figure 7. Example of one pathway for determining Creditability using the Decision Tree

Additionally, we calculated the score of the Predicted Value vs the Actual Value with results of 0.70 or 70%. This states that the model correctly predicted the outcome 70% of the time.

Below is the prediction on the testing data taking into consideration all attributes from the training data. The decision tree model has predicted that the first record is '0' for Creditability which means bad credit. The second record shows Credibility as '1' which is good credit.

```
#unfiltered dataset
# Let's make the prdictions on the test set  that we set aside earlier using the trained tree
y_pred = clf.predict(X_test)
y_pred
```

```
array(['0', '1', '1', '1', '1', '0', '1', '0', '1', '0', '1', '1', '0',
       '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '1', '1',
       '1', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '0',
       '0', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '1',
       '0', '1', '1', '1', '0', '1', '1', '1', '1', '1', '1', '1', '0',
       '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1',
       '1', '1', '1', '1', '0', '0', '1', '1', '1', '1', '1', '1', '1',
       '1', '1', '1', '0', '1', '1', '1', '1', '1', '0', '0', '1', '1',
       '1', '0', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '0',
       '1', '0', '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1',
       '1', '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '1', '0',
       '0', '1', '1', '0', '1', '1', '1', '1', '1', '1', '1', '1', '0',
       '1', '1', '1', '1', '1', '1', '0', '1', '1', '1', '1', '1', '1',
       '1', '1', '1', '1', '1', '1', '0', '1', '0', '0', '1', '1', '1',
       '0', '1', '1', '1', '1', '1', '1', '1', '0', '0', '0', '1', '1',
       '1', '0', '1', '0', '1'], dtype=object)
```

Figure 6. Array of prediction on a test set.

While performing the decision tree algorithm, the response for our prediction as described in the confusion matrix is summarized below:

Out of 200 tests, the performance measures for the prediction data are as follows:

Confusion Matrix	**Predicted 0**	**Predicted 1**
Actual 0	TN: 17	FP: 39
Actual 1	FN: 20	TP: 124

Table 5. Confusion Matrix on decision tree (original data set).

True-Positives (correctly predicted) successfully 124 times. False-Negatives (wrongly predicted) 20 times. False-Negatives are the most harmful in our case, also known as type II errors where the bank would lose out on an applicant's business if they are incorrectly classified as having bad credit when in fact, the applicant has good credit. False-Positives are type I errors where applicants are incorrectly deemed to have good credit when they don't.

The precision, recall, and accuracy score would be as follows:

	Precision	**Recall**	**F1 Score**	**Support**
0	**0.46**	**0.30**	**0.70**	56
1	**0.76**	**0.88**	**0.81**	144
Accuracy			**0.70**	200
macro avg	0.61	0.58	0.59	200
weighted avg	0.68	0.70	0.68	200

Table 6. Precision, recall, and accuracy scores for decision tree (original data set).

Additional Testing:

To further test our predicted data set, we can create a mock-applicant data set similar to the first row of data to check the outcome of the predicted data. We can then compare to determine our how confident are in our algorithm:

1. Account Balance: 1
2. Duration of Credit (month): 3
3. Payment Status of Previous Credit: 3
4. Purpose: 9
5. Credit Amount: 5868
6. Value Savings/Stocks: 2
7. Length of current employment:4
8. Installment percent: 1
9. Sex & Marital Status: 4
10. Guarantors: 1
11. Duration in Current address: 4
12. Most valuable available asset: 1
13. Age (years): 30
14. Concurrent Credits: 1
15. Type of apartment: 2
16. No of Credits at this Bank: 1
17. Occupation: 4
18. No of dependents: 1
19. Telephone: 1

20. Foreign Worker: 1

After running our mock-applicant through our predictive model and comparing with the predicted data set, the results showed us that our mock-applicant has good credibility, thus supporting the confidence of our predictive algorithm.

Naive Bayes

Naive Bayes (NB) is a classification technique based on the Bayes Theorem with the assumption of independence among predictors. In using NB, we are assuming that the presence of all the attributes are unrelated with any other attributes as predictors for good or bad credit.

Multinomial Naive Bayes algorithm was used over the 20 attributes and 1 class attribute, Creditability. Using the same train-test split set strategy, the summary for the Naive Bayes classification is as follows:

Number of features used is 20 (1 class: Creditability)
Number of records for class 0: 244
Number of records for class 1: 556
Log probability*: -1.1874435 for class 0
Log probability*: -0.36384343 for class 1

**Log conditional probability for each feature given a class means that it calculated the probability of getting a particular class by each attribute.*

The precision, recall, and accuracy score would be as follows:

	Precision	Recall	F1 Score	Support
0	**0.35**	**0.36**	**0.35**	56
1	**0.75**	**0.74**	**0.75**	144
Accuracy			**0.63**	200
macro avg	0.55	0.55	0.55	200
weighted avg	0.64	0.64	0.64	200

Table 8. Precision, recall, and accuracy scores on Naive Bayes (original data set).

Summary of Predictive Modelling

For predicting class '1'	Decision Tree	Naive Bayes
Accuracy	0.70	0.63
Precision	0.76	0.75

Recall	0.86	0.74
F1 Score	0.81	0.75

Table 9. Comparison of performance of the classifiers.

Overall, using the decision tree algorithm performed the best of the two classifiers. The accuracy is higher in the decision tree at 70% while with Naive Bayes algorithm, we only were able to get 64% accuracy. The decision tree algorithm also outperformed in comparison in the other evaluation metrics in precision and recall which results in an overall higher F1-score.

Post Predictive Analysis

During data preparation, we were able to identify 4 attributes with visual and statistical analyses that determined low or no correlation between them and the target variable. Using the ANOVA and Chi-Square tests to determine correlation, we returned the following 4 attributes to filter out from the original data set:

1. Installment percent
2. Duration in Current address
3. No of Credits at this Bank
4. No of dependents

The new filtered data set will contain 17 variables (inclusive of the target variable) and to confirm our correlation analyses, we will rerun the new filtered data set through both decision tree and Naive Bayes algorithms.

For predicting class '1'	**Decision Tree** Original data set	**Decision Tree** Filtered data set	**Naive Bayes** Original data set	**Naive Bayes** Filtered data set
Accuracy	0.70	**0.74**	0.63	**0.64**
Precision	0.76	**0.80**	0.75	**0.75**
Recall	0.86	**0.86**	0.74	**0.74**
F1 Score	0.81	**0.83**	0.75	**0.75**

Table 10. Comparison of performance of the classifiers between the original and filtered data sets.

Summary of Post Predictive Analysis

We can conclude that the results from rerunning the algorithm with the filtered data set shows an increase in overall score with the decision tree classifier. The accuracy rate increased from 70% to 74%, and the precision score increased 4%. Comparing the results of the Naive Bayes classifier, we see a very slight increase in accuracy from 63% to 64%, but the other evaluation

metrics stay the same.

Conclusions and Recommendations

After our modeling and classifying, we would recommend to the client bank that the key factors to grant a client credit are “Account Balance”, “Duration of Credit”, “Credit Amount”, “Payment Status of Previous Credit”, and “Purpose of Credit”. This information will be very helpful for the client to determine whether the person is credible.

Through the different methods of modeling, we know that the filtered data gives us better accuracy. However, in real life, the dataset would be a lot larger, therefore, we require more testing to validate whether these are only factors. The dataset might neglect some other information that would be helpful to validate the clients. For example, it might be helpful to know the client’s income (the whole family), or credit score. Furthermore, it is possible with different cultures or varieties that will differ in other countries if the client bank decides to expand.