German Credit Card Analysis

Group 3

Gaurang Dadu, gdadu@torontomu.ca
Anjali Verma, a10verma@torontomu.ca

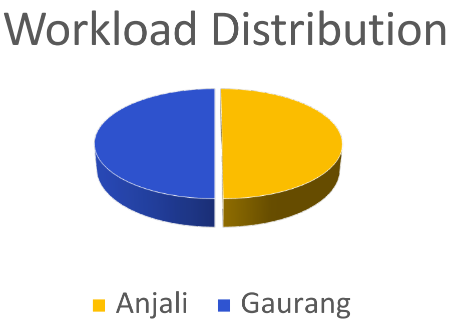
RYERSON
UNIVERSITY

Introduction

- The aim of this project is to develop a data analytics strategy for loan approval decisions using German Credit Data's 20 attributes and a class attribute with a dataset of 1000 rows.
- Developed a data analytics-based strategy that can assist bank managers in making credit approval decisions by identifying most effective features and models.
- Data preparation and modelling were performed using Python, while SAS and Python’s Weka module were used to help determine the selected attributes.
- The project employed Decision Tree and Naïve Bayes algorithms, using Supervised Learning Classification method in Machine Learning, to analyze all and selected attributes.
- Naïve Bayes using selected attributes achieved the highest accuracy at 76% while minimizing false positives in the generated model.

Workload Distribution

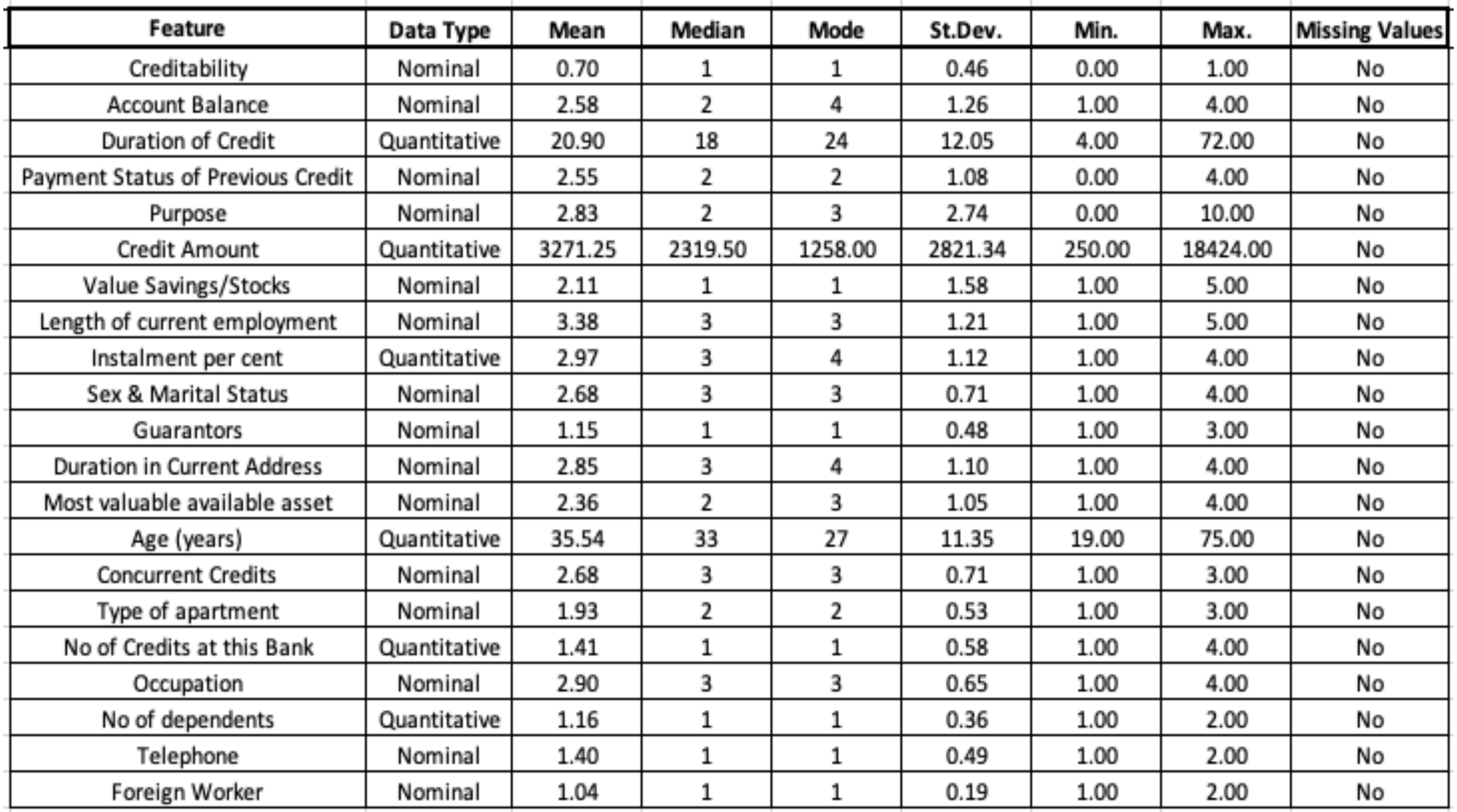
Member Name	List of Tasks Performed
Gaurang Dadu	Introduction, Data Preparation, Predictive Modelling, Conclusion & Recommendation, Presentation
Anjali Verma	Introduction, Data Preparation, Predictive Modelling, Conclusion & Recommendation, Presentation



Data Preparation

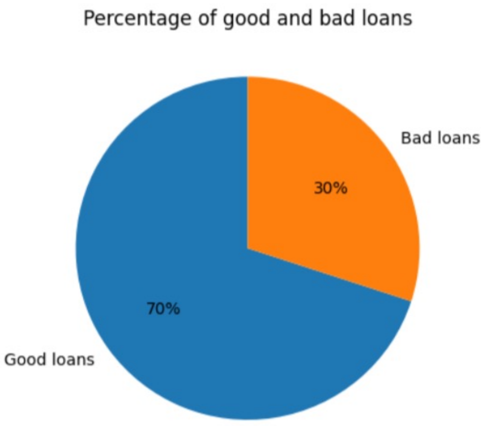
- Checked types of all 21 attributes: 6 numerical, 15 categorical/nominal. "Creditability" is our class attribute which is categorical.
- No missing values found in the dataset.
- Obtained Max, Min, and Standard deviation of attributes to ensure data quality and validity.
- Checked percentage of good and bad loans to evaluate data analytics strategy and model effectiveness.
- Plotted histograms to visualize data distribution and found data to be imbalanced.
- Correlated each attribute with the class attribute and used SAS and Weka to get suggestions for select attributes.
- Found high positive correlation between "Duration of Credit (month)" and "Credit Amount" and decided to consider only one to avoid redundancy.
- Finally selected these features: "Duration of Credit (month)", "Account Balance", "Payment Status of Previous Credit" and "Value Savings/Stock".

Data Preparation: Descriptive Statistics

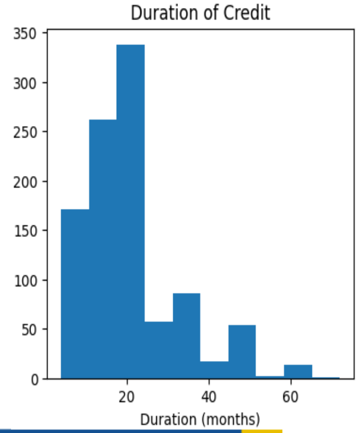


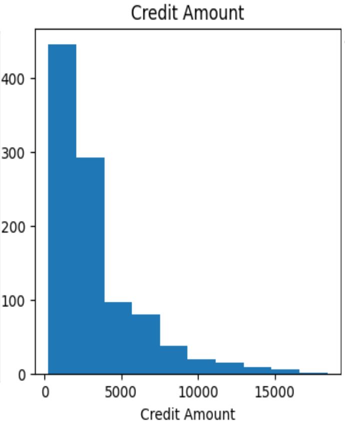
Feature	Data Type	Mean	Median	Mode	St.Dev.	Min.	Max.	Missing Values
Creditability	Nominal	0.70	1	1	0.46	0.00	1.00	No
Account Balance	Nominal	2.58	2	4	1.26	1.00	4.00	No
Duration of Credit	Quantitative	20.90	18	24	12.05	4.00	72.00	No
Payment Status of Previous Credit	Nominal	2.55	2	2	1.08	0.00	4.00	No
Purpose	Nominal	2.83	2	3	2.74	0.00	10.00	No
Credit Amount	Quantitative	3271.25	2319.50	1258.00	2821.34	250.00	18424.00	No
Value Savings/Stocks	Nominal	2.11	1	1	1.58	1.00	5.00	No
Length of current employment	Nominal	3.38	3	3	1.21	1.00	5.00	No
Instalment per cent	Quantitative	2.97	3	4	1.12	1.00	4.00	No
Sex & Marital Status	Nominal	2.68	3	3	0.71	1.00	4.00	No
Guarantors	Nominal	1.15	1	1	0.48	1.00	3.00	No
Duration in Current Address	Nominal	2.85	3	4	1.10	1.00	4.00	No
Most valuable available asset	Nominal	2.36	2	3	1.05	1.00	4.00	No
Age (years)	Quantitative	35.54	33	27	11.35	19.00	75.00	No
Concurrent Credits	Nominal	2.68	3	3	0.71	1.00	3.00	No
Type of apartment	Nominal	1.93	2	2	0.53	1.00	3.00	No
No of Credits at this Bank	Quantitative	1.41	1	1	0.58	1.00	4.00	No
Occupation	Nominal	2.90	3	3	0.65	1.00	4.00	No
No of dependents	Quantitative	1.16	1	1	0.36	1.00	2.00	No
Telephone	Nominal	1.40	1	1	0.49	1.00	2.00	No
Foreign Worker	Nominal	1.04	1	1	0.19	1.00	2.00	No

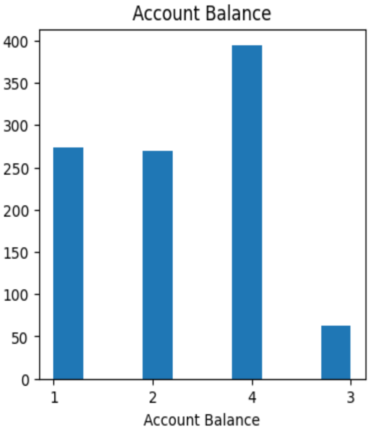
Data Preparation –Visualization (Count as per class)

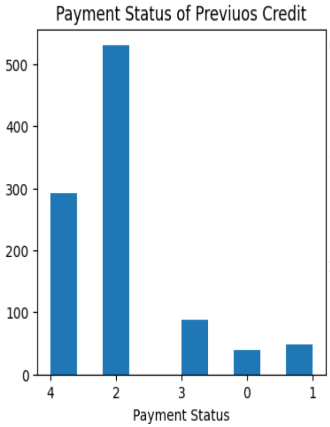


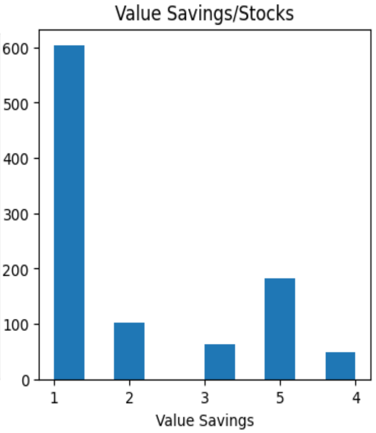
Feature	Correlation coefficient with class attribute
Duration of Credit	0.62
Credit Amount	0.62
Account Balance	0.35
Payment status of previous credit	0.23
Value of Stock	0.18



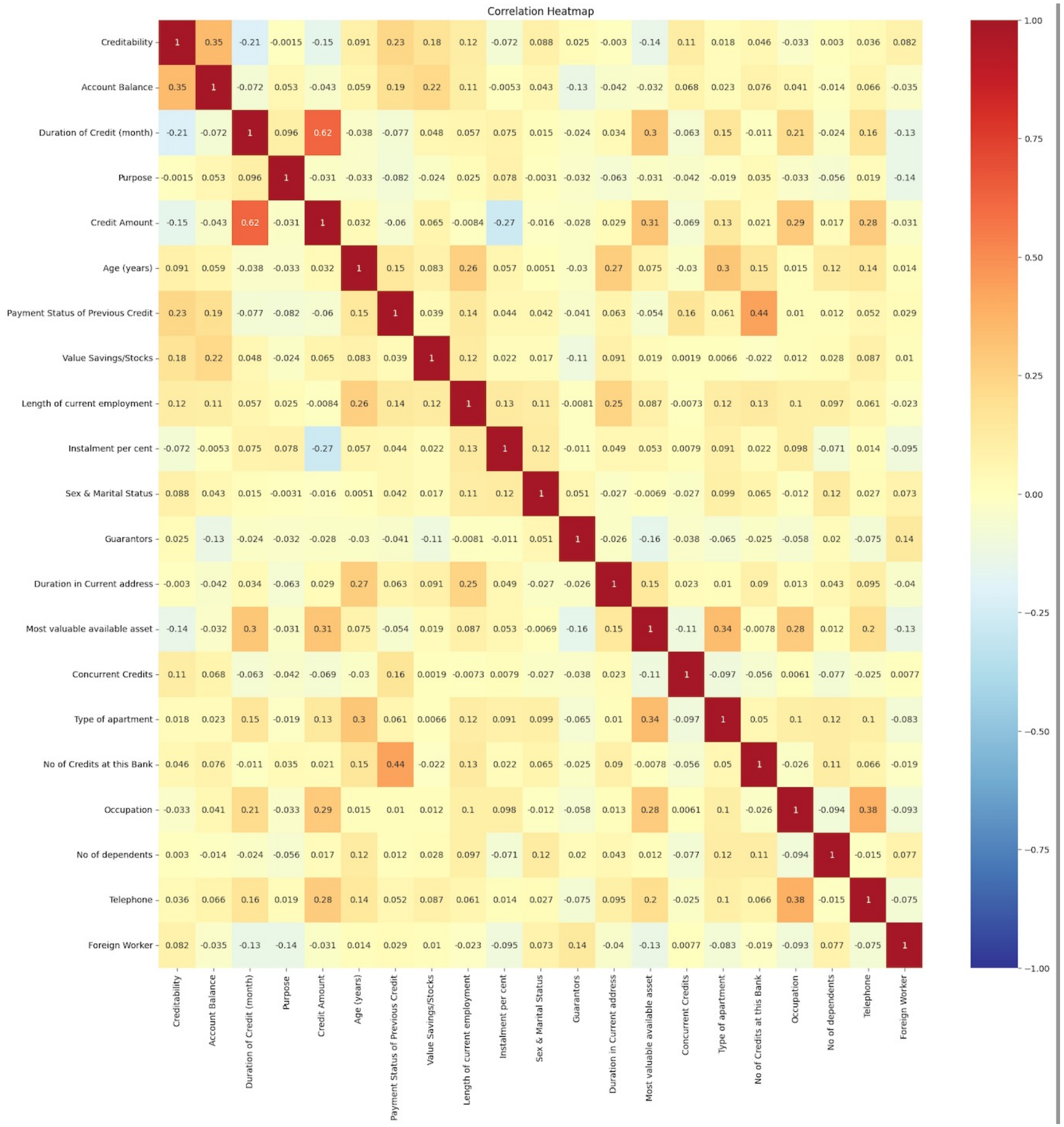








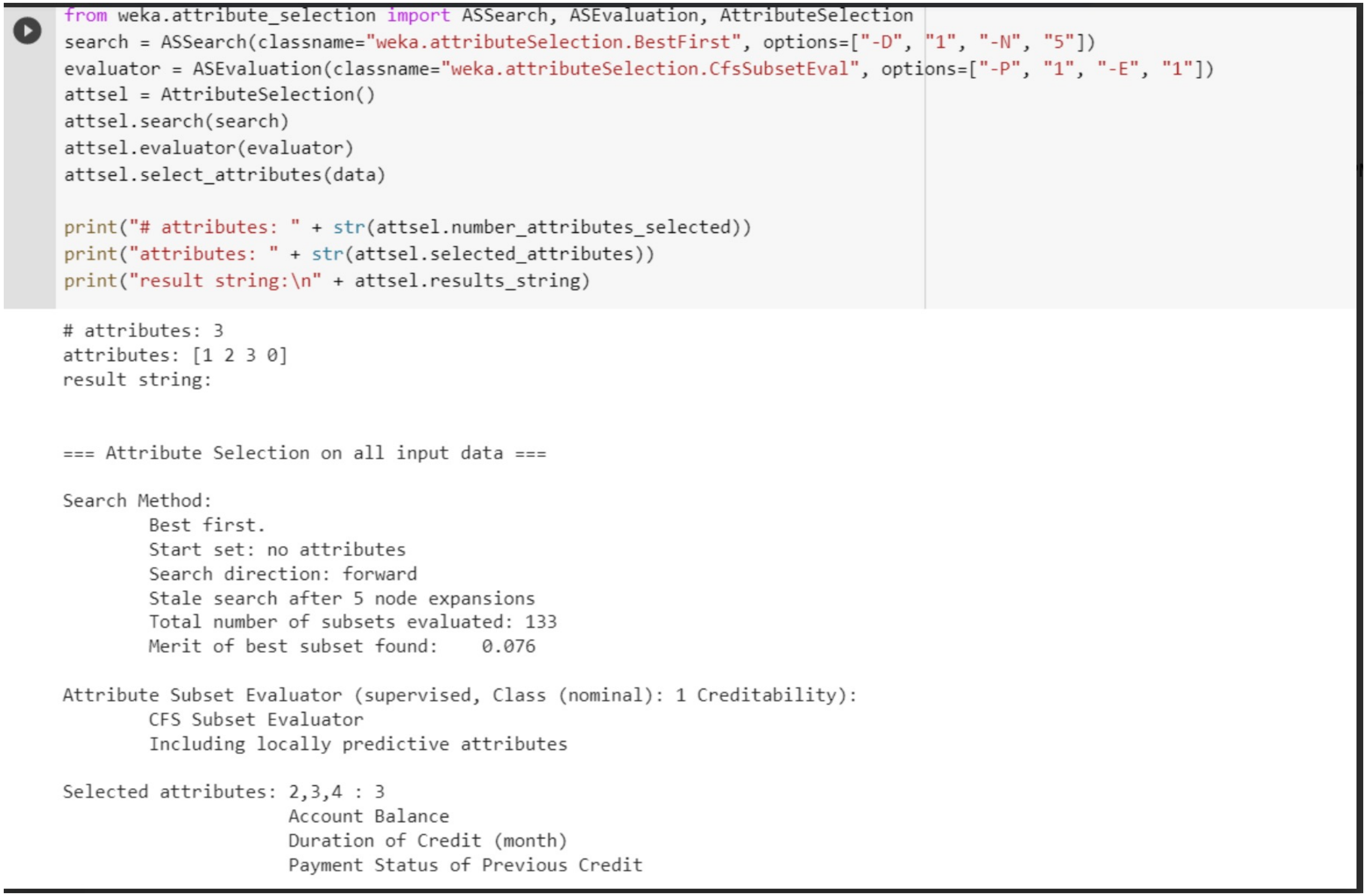
Correlation Heatmap of each attribute with the class attribute



Correlation Heatmap

	Creditability	Account Balance	Duration of Credit (month)	Purpose	Credit Amount	Age (years)	Payment Status of Previous Credit	Value Savings/Stocks	Length of current employment	Instalment per cent	Sex & Marital Status	Guarantors	Duration in Current address	Most valuable available asset	Concurrent Credits	Type of apartment	No of Credits at this Bank	Occupation	No of dependents	Telephone	Foreign Worker
Creditability	1	0.35	-0.21	-0.0015	-0.15	0.091	0.23	0.18	0.12	-0.072	0.088	0.025	-0.003	-0.14	0.11	0.018	0.046	-0.033	0.003	0.036	0.082
Account Balance	0.35	1	-0.072	0.053	-0.043	0.059	0.19	0.22	0.11	-0.0053	0.043	-0.13	-0.042	-0.032	0.068	0.023	0.076	0.041	-0.014	0.066	-0.035
Duration of Credit (month)	-0.21	-0.072	1	0.096	0.62	-0.038	-0.077	0.048	0.057	0.075	0.015	-0.024	0.034	0.3	-0.063	0.15	-0.011	0.21	-0.024	0.16	-0.13
Purpose	-0.0015	0.053	0.096	1	-0.031	-0.033	-0.082	-0.024	0.025	0.078	-0.0031	-0.032	-0.063	-0.031	-0.042	-0.019	0.035	-0.033	-0.056	0.019	-0.14
Credit Amount	-0.15	-0.043	0.62	-0.031	1	0.032	-0.06	0.065	-0.0084	-0.27	-0.016	-0.028	0.029	0.31	-0.069	0.13	0.021	0.29	0.017	0.28	-0.031
Age (years)	0.091	0.059	-0.038	-0.033	0.032	1	0.15	0.083	0.26	0.057	0.0051	-0.03	0.27	0.075	-0.03	0.3	0.15	0.015	0.12	0.14	0.014
Payment Status of Previous Credit	0.23	0.19	-0.077	-0.082	-0.06	0.15	1	0.039	0.14	0.044	0.042	-0.041	0.063	-0.054	0.16	0.061	0.44	0.01	0.012	0.052	0.029
Value Savings/Stocks	0.18	0.22	0.048	-0.024	0.065	0.083	0.039	1	0.12	0.022	0.017	-0.11	0.091	0.019	0.0019	0.0066	-0.022	0.012	0.028	0.087	0.01
Length of current employment	0.12	0.11	0.057	0.025	-0.0084	0.26	0.14	0.12	1	0.13	0.11	-0.0081	0.25	0.087	-0.0073	0.12	0.13	0.1	0.097	0.061	-0.023
Instalment per cent	-0.072	-0.0053	0.075	0.078	-0.27	0.057	0.044	0.022	0.13	1	0.12	-0.011	0.049	0.053	0.0079	0.091	0.022	0.098	-0.071	0.014	-0.095
Sex & Marital Status	0.088	0.043	0.015	-0.0031	-0.016	0.0051	0.042	0.017	0.11	0.12	1	0.051	-0.027	-0.0069	-0.027	0.099	0.065	-0.012	0.12	0.027	0.073
Guarantors	0.025	-0.13	-0.024	-0.032	-0.028	-0.03	-0.041	-0.11	-0.0081	-0.011	0.051	1	-0.026	-0.16	-0.038	-0.065	-0.025	-0.058	0.02	-0.075	0.14
Duration in Current address	-0.003	-0.042	0.034	-0.063	0.029	0.27	0.063	0.091	0.25	0.049	-0.027	-0.026	1	0.15	0.023	0.01	0.09	0.013	0.043	0.095	-0.04
Most valuable available asset	-0.14	-0.032	0.3	-0.031	0.31	0.075	-0.054	0.019	0.087	0.053	-0.0069	-0.16	0.15	1	-0.11	0.34	-0.0078	0.28	0.012	0.2	-0.13
Concurrent Credits	0.11	0.068	-0.063	-0.042	-0.069	-0.03	0.16	0.0019	-0.0073	0.0079	-0.027	-0.038	0.023	-0.11	1	-0.097	-0.056	0.0061	-0.077	-0.025	0.0077
Type of apartment	0.018	0.023	0.15	-0.019	0.13	0.3	0.061	0.0066	0.12	0.091	0.099	-0.065	0.01	0.34	-0.097	1	0.05	0.1	0.12	0.1	-0.083
No of Credits at this Bank	0.046	0.076	-0.011	0.035	0.021	0.15	0.44	-0.022	0.13	0.022	0.065	-0.025	0.09	-0.0078	-0.056	0.05	1	-0.026	0.11	0.066	-0.019
Occupation	-0.033	0.041	0.21	-0.033	0.29	0.015	0.01	0.012	0.1	0.098	-0.012	-0.058	0.013	0.28	0.0061	0.1	-0.026	1	-0.094	0.38	-0.093
No of dependents	0.003	-0.014	-0.024	-0.056	0.017	0.12	0.012	0.028	0.097	-0.071	0.12	0.02	0.043	0.012	-0.077	0.12	0.11	-0.094	1	-0.015	0.077
Telephone	0.036	0.066	0.16	0.019	0.28	0.14	0.052	0.087	0.061	0.014	0.027	-0.075	0.095	0.2	-0.025	0.1	0.066	0.38	-0.015	1	-0.075
Foreign Worker	0.082	-0.035	-0.13	-0.14	-0.031	0.014	0.029	0.01	-0.023	-0.095	0.073	0.14	-0.04	-0.13	0.0077	-0.083	-0.019	-0.093	0.077	-0.075	1

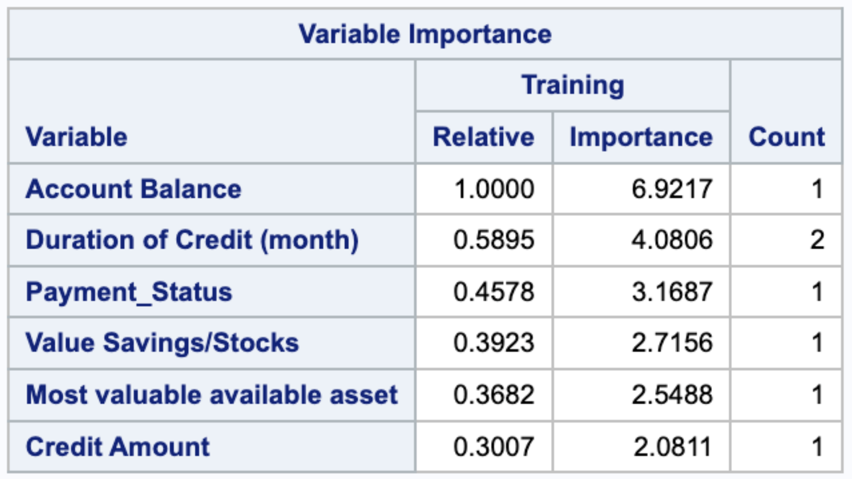
The following code shows the use of Python’s Weka module to get Suggestions to determine selected attributes.



```
from weka.attribute_selection import ASSearch, ASEvaluation, AttributeSelection
search = ASSearch(classname="weka.attributeSelection.BestFirst", options=["-D", "1", "-N", "5"])
evaluator = ASEvaluation(classname="weka.attributeSelection.CfsSubsetEval", options=["-P", "1", "-E", "1"])
attsel = AttributeSelection()
attsel.search(search)
attsel.evaluator(evaluator)
attsel.select_attributes(data)

print("# attributes: " + str(attsel.number_attributes_selected))
print("attributes: " + str(attsel.selected_attributes))
print("result string:\n" + attsel.results_string)
```

```
# attributes: 3
attributes: [1 2 3 0]
result string:


=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 133
        Merit of best subset found:    0.076

Attribute Subset Evaluator (supervised, Class (nominal): 1 Creditability):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,4 : 3
                     Account Balance
                     Duration of Credit (month)
                     Payment Status of Previous Credit
```

The following image shows the output from SAS. It gives us the important variables in our dataset.
We used SAS to get suggestions for determining selected attributes.



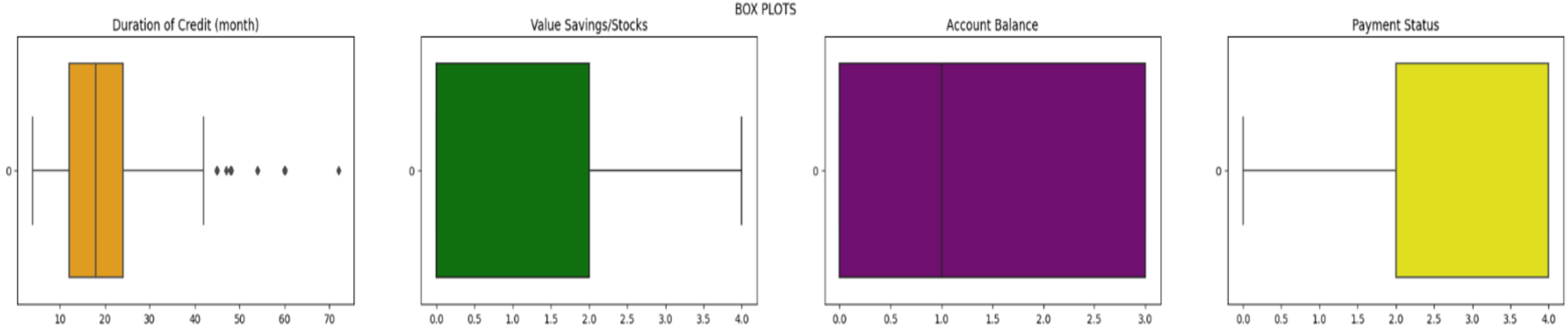
Variable Importance

	Training		
Variable	Relative	Importance	Count
Account Balance	1.0000	6.9217	1
Duration of Credit (month)	0.5895	4.0806	2
Payment_Status	0.4578	3.1687	1
Value Savings/Stocks	0.3923	2.7156	1
Most valuable available asset	0.3682	2.5488	1
Credit Amount	0.3007	2.0811	1

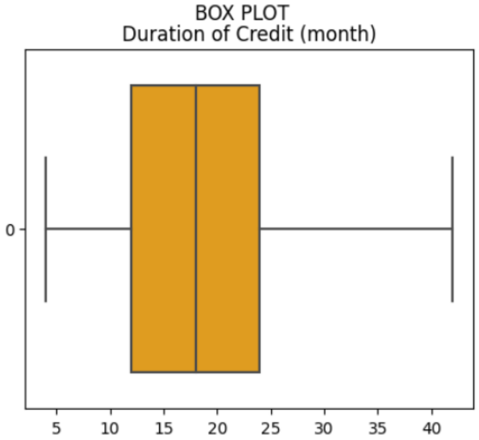
After plotting the Correlation Heatmap and getting suggestions from
Weka and SAS. We decided to select these features for our predictive modelling-

- Duration of Credit (month)
- Account Balance
- Payment Status of Previous Credit
- Value Savings/Stocks

- From the selected attributes, the categorical attributes were converted to numeric and were checked for outliers by printing box plots. The only attribute that had outliers was “Duration of Credit”.



- Removed outliers from the ”Duration of Credit” attribute to ensure quality and validity of the data. However, due to small dataset size, chose to use the original data without removing the outliers for further predictive modelling.



Predictive Modeling

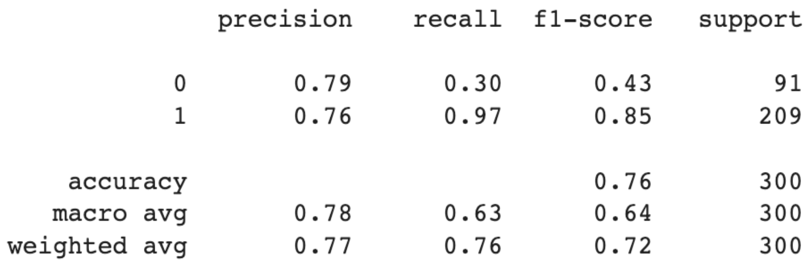
- **Decision Tree and Naïve Bayes classification algorithms were used to predict the class attribute on both, baseline (all attributes) and selected attributes.**
- **The dataset was split into training and test sets using the train-test split method with a 70:30 ratio.**
- **Train-test split method was used because of it's ability to generalize new data and to evaluate the performance of the model on new, unseen data and avoid overfitting.**
- **The performance of the algorithms was measured using true positive, false positive and accuracy measures.**
- **The algorithm with best performance was determined based on the performance measures obtained from the output of both methods.**

Predictive Modeling

Results

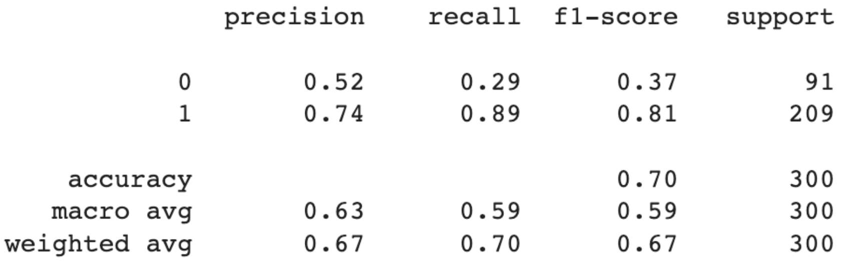
Model	Accuracy	True Positive	False Positive
Baseline decision tree	75%	202 (67%)	65 (22%)
Baseline Naïve Bayes	62%	149 (50%)	54 (18%)
Decision tree on selected features	70%	185 (62%)	65 (22%)
Naïve Bayes on selected features	76%	172 (57%)	36 (12%)

Here we can see the Precision, Recall and f1-score for all of the Predictive models that were used.



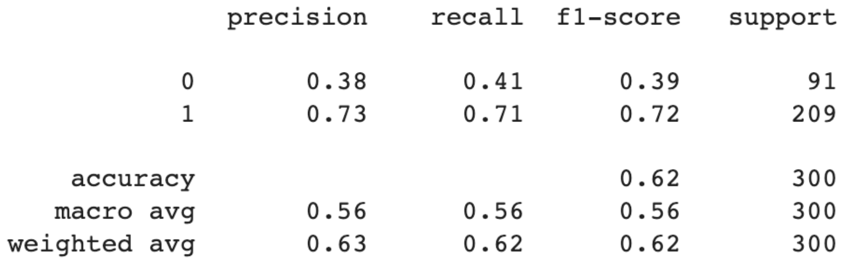
```
              precision    recall  f1-score   support

           0       0.79      0.30      0.43        91
           1       0.76      0.97      0.85       209

    accuracy                           0.76       300
   macro avg       0.78      0.63      0.64       300
weighted avg       0.77      0.76      0.72       300
```

Baseline Decision Tree



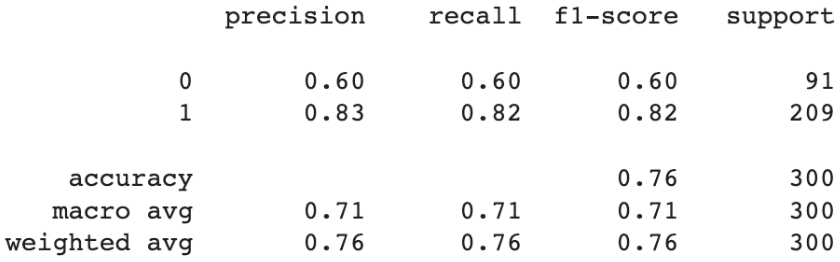
```
              precision    recall  f1-score   support

           0       0.52      0.29      0.37        91
           1       0.74      0.89      0.81       209

    accuracy                           0.70       300
   macro avg       0.63      0.59      0.59       300
weighted avg       0.67      0.70      0.67       300
```

Decision Tree on Selected Features



```
              precision    recall  f1-score   support

           0       0.38      0.41      0.39        91
           1       0.73      0.71      0.72       209

    accuracy                           0.62       300
   macro avg       0.56      0.56      0.56       300
weighted avg       0.63      0.62      0.62       300
```

Baseline Naïve Bayes



```
              precision    recall  f1-score   support

           0       0.60      0.60      0.60        91
           1       0.83      0.82      0.82       209

    accuracy                           0.76       300
   macro avg       0.71      0.71      0.71       300
weighted avg       0.76      0.76      0.76       300
```

Naïve Bayes on Selected Features

Predictive Modeling

Decision Tree

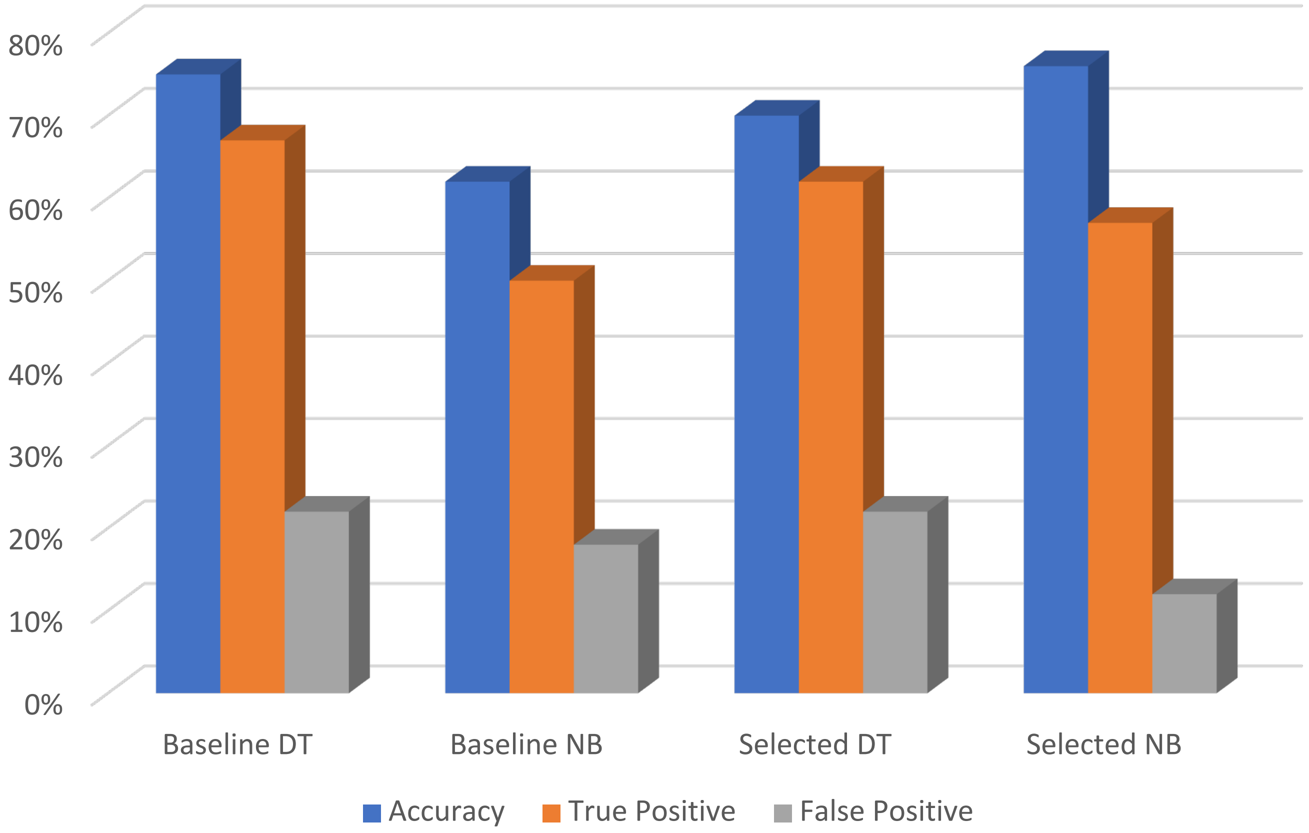
- **Two decision trees were made: one with all attributes(baseline tree) and the other with only selected attributes.**
- **The accuracy of baseline tree is 75%, with true positive and false positive values of 67% and 22% respectively.**
- **The accuracy of selected features tree is 70%, with true positive and false positive values of 62% and 22% respectively.**
- **The comparison of the two decision trees suggests that the baseline model, which uses all attributes, is better for identifying positive cases as it has higher accuracy and true positive rate. The selected features model has a lower accuracy and true positive rate.**

Predictive Modeling

Naïve Bayes

- **Two Naïve Bayes models were created to evaluate the performance: the baseline model with all attributes and the selected attributes model.**
- **The baseline model had an accuracy of 62%, with true positive and false positive values of 50% and 18% respectively.**
- **The selected attributes model had an accuracy of 76%, with true positive and false positive values of 57% and 12% respectively.**
- **The performance of the Naïve Bayes model improved significantly with the use of selected features, as the accuracy increased from 62% to 76% and the false positive rate decreased from 18% to 12%.**

Predictive Model Comparison



Conclusion and Recommendation

- **The Decision Tree model out performed Naïve Bayes in terms of accuracy, but Naïve Bayes had a lower false positive rate, indicating its ability to identify low-risk borrowers.**
- **Based on the evaluation, we recommend using Naïve Bayes with selected features for future predictive modeling, as it significantly improves accuracy.**
- **However, both models have relatively low accuracy, indicating the need for additional features to improve performance.**
- **We recommend adding more features like Credit Score and Income to the dataset to improve accuracy.**
- **Some features, such as No. of dependents, Telephone and Duration in Current address have a negative correlation with the class feature, and removing them would simplify the dataset and improve usability.**