LLK Intelligence

Predicting Credit-worthy Applicants

Aidan Kowolik
aidan.kowolik@ryerson.ca

Matthew Lee
matthew2.lee@ryerson.ca

Kai Lu
k10lu@ryerson.ca

Meet our team

Aidan

- Data Exploration and Preparation

Matthew

- Predictive Modelling:
  - Decision Tree

Kai

- Predictive Modelling
  - Naive Bayes
- Further exploratory Analysis
  - Linear regression
  - Correlation matrix

Team

Developed analysis and presentation materials

LLK Intelligence

Introduction

Problem statement

Which credit customers should get approval for a loan?

Our task

Recommend a strategy based on available credit data that will help bank managers decide whether to approve a loan for new applicants

Our conclusion

WEKA-based Naive Bayes model based four attributes performs the best, while satisfying model objectives

Data Exploration and Preparation

STEP 1

Check for missing data

STEP 2

Complete distributional analysis

STEP 3

Detect notable data outliers

LLK Intelligence

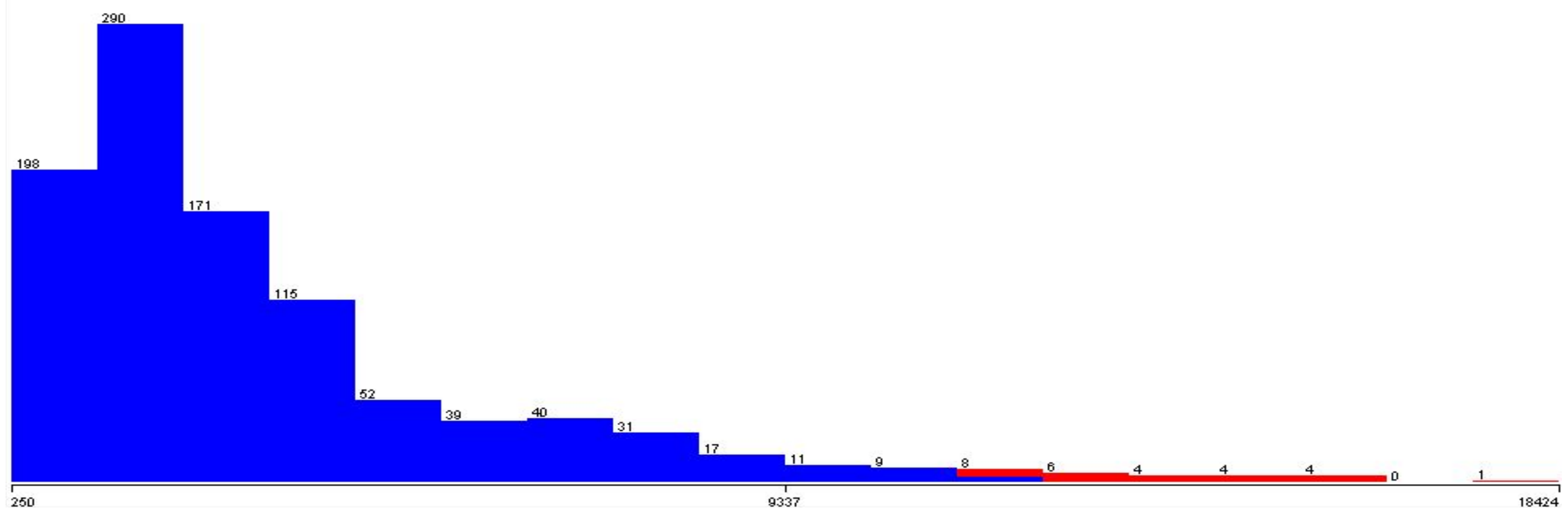
Data Preparation Highlights

Quartile Distribution of Most Relevant Fields

	Duration of Credit (mos)	Credit Amount
Min	4	250
25th Percentile	12	1,366
Median	18	2,320
75th Percentile	24	3,972
Max	72	18,424

Credit Amount Distribution

Outliers removed in the prepared dataset



- Most fields in dataset are qualitative classes despite quantitative in nature
- Focused on preparing data for ‘`numeric`’ fields
- Notable outliers in `Credit Amount` and removed those records in the prepared dataset

LLK Intelligence

Data Analysis

LLK Intelligence

Predictive Modelling Approach

CONSERVATIVE APPROACH

Keep money safe first, then profit
Assume a traditional loan business that is risk averse

PRECISION MOST IMPORTANT

Most important:
- to correctly identify clients with good credit
- to minimize approving applicants with bad credit

		PREDICTED	
		Positive *Good Creditability*	**Negative** *Bad Creditability*
ACTUAL	**Positive** *Good Creditability*	True Positive	False Negative
	Negative *Bad Creditability*	False Positive	True Negative

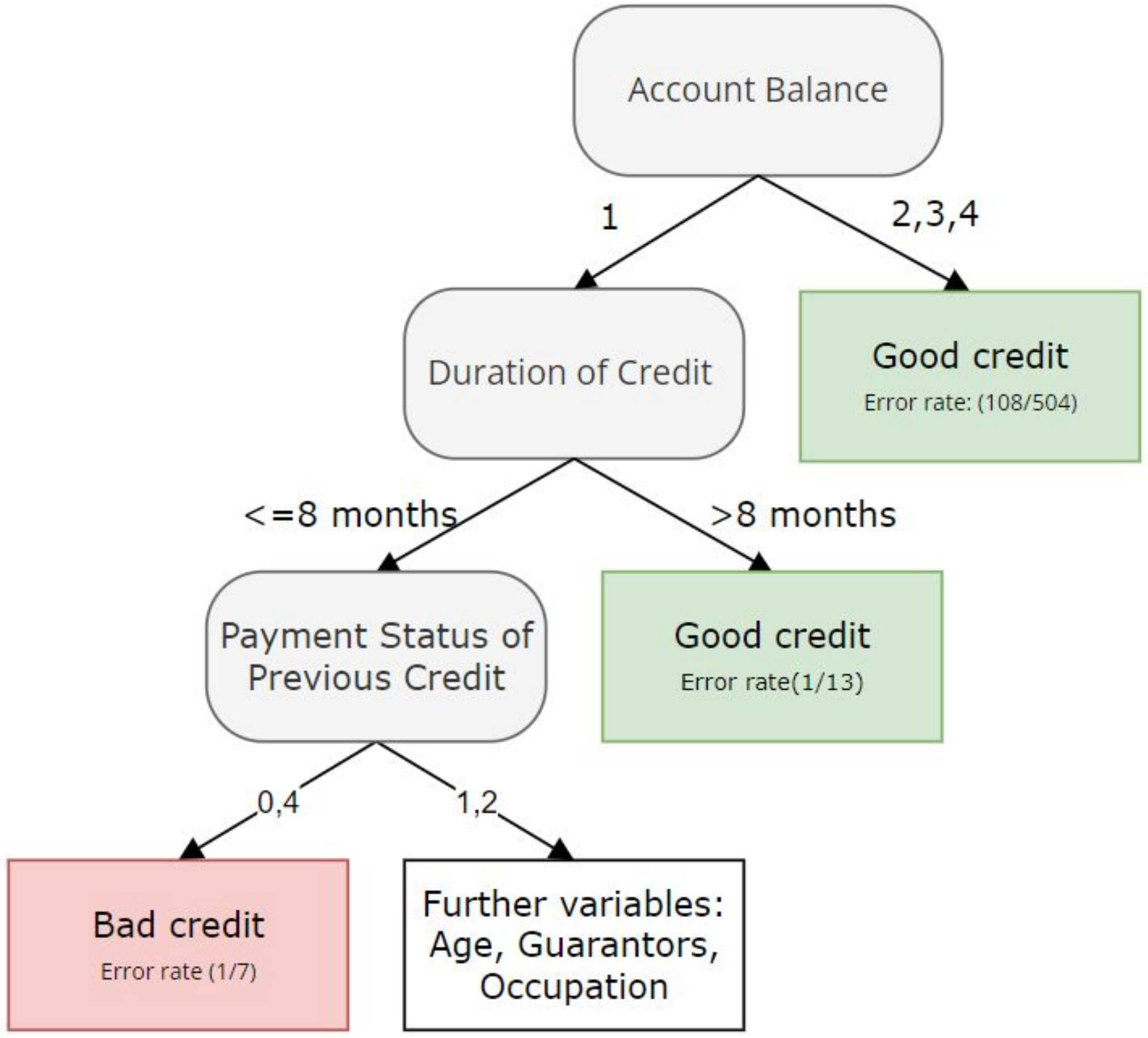
MOST IMPORTANT

MINIMIZE DATA FIELDS IN MODEL

... but only if it does not affect precision
Limit the amount of fields in the model will allow for faster response times to applicants

Predictive Modelling Methods

Decision Tree Output Example



Decision Tree Method

- Predicts the value of a target variable by learning simple decision rules inferred from the data features
- Key variables that predict good credit applicants:
  - Account balance, Value of Savings / Stocks, Payment Status of Previous Credit, Duration of Credit, Age, Guarantors

Naive Bayes Method

- Most important conditional probabilities:
  - Account balance, Payment Status of Previous Credit, Duration of Credit

LLK Intelligence

Predictive Modelling Results

	Precision (Most Important)	Accuracy	Recall
Pre-prepared data - Baseline [n = 300]			
Decision Tree	0.70	0.65	0.80
Naive Bayes	0.76	0.75	0.90
Prepared data [n = 293]			
Decision Tree	0.82	0.75	0.81
Naive Bayes	0.84	0.76	0.81

Removal of outliers led to improved Precision figures

Naive Bayes performed marginally better than Decision Tree and process to next step

LLK Intelligence

Predictive Modelling Refinement

- Recommend the Naive Bayes classification method
- Selected the commonly recurring field in all classification models and re-ran the model

1 Account Balance

2 Payment Status of Previous Credit

3 Duration of Credit

4 Value Savings and Stock

LLK Intelligence

Predictive Modelling Refinement

Naive Bayes Model

	Precision (Most Important)	Accuracy	Recall
Pre-prepared data Baseline [n = 300]	0.76	0.75	0.90
Prepared data All variables [n = 293]	0.84	0.76	0.81
Prepared data Four variables [n = 293]	0.82	0.76	0.84

Marginal impact to Precision by isolating to only four variables

Simplifying the model is worthwhile to reduce complexity and improve decision response times

LLK Intelligence

Recommendations

Apply the Naive Bayes classification model based on the following criteria:

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

Require further training of the model

Seek expanded dataset with more numeric fields, allowing the model to identify its own classification ‘bins’

Appendices

Predictive Model Development and Selection

Run Decision Tree and Naive Bayes classifiers using the WEKA application

1

Run based on a ratio 70/30 train and test sets on the raw and prepared datasets

2

Choose the classification model and isolate the model to smaller set of variables where possible

3

LLK Intelligence

Metrics selection logic

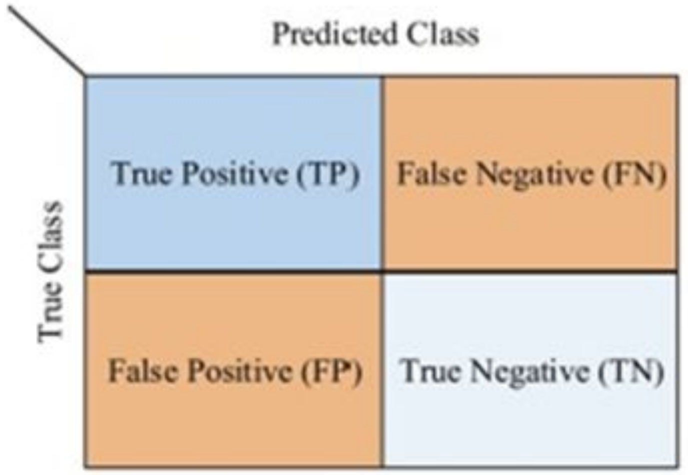
bank

Money safe first, then profit

In all predict 1(+) datapoints, the number of actual 1(+) datapoints should as large as possible

Precision should as high as possible

We can tolerate a certain amount of FN, but we have zero tolerance for FP.
Cause FP means we provide loans to class 0 and then we cannot get money back.

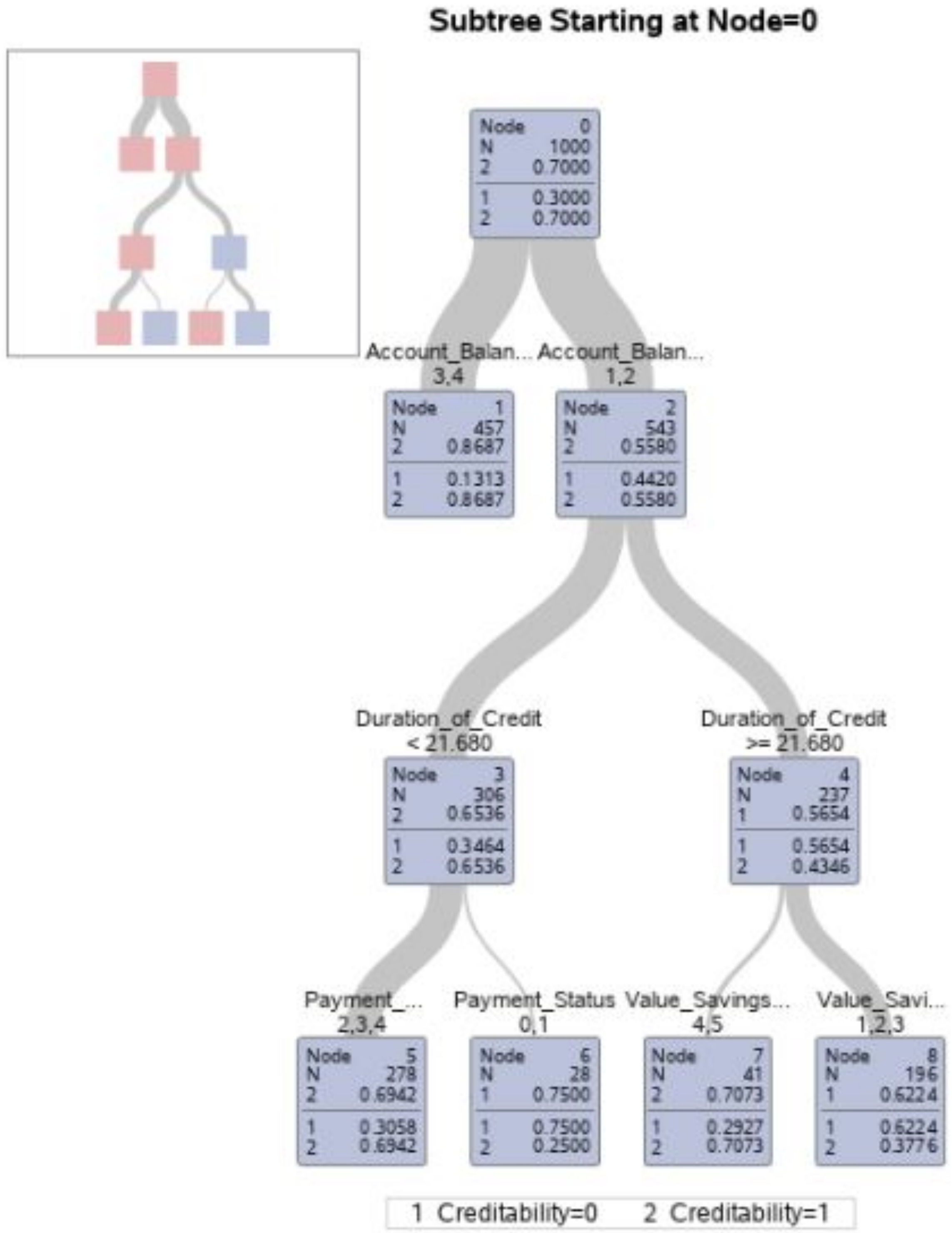


We assume the bank belongs to the risk-averse type (cause just the traditional loan business, not risk-seeking type like doing hedge fund), So Precision is our main consideration.

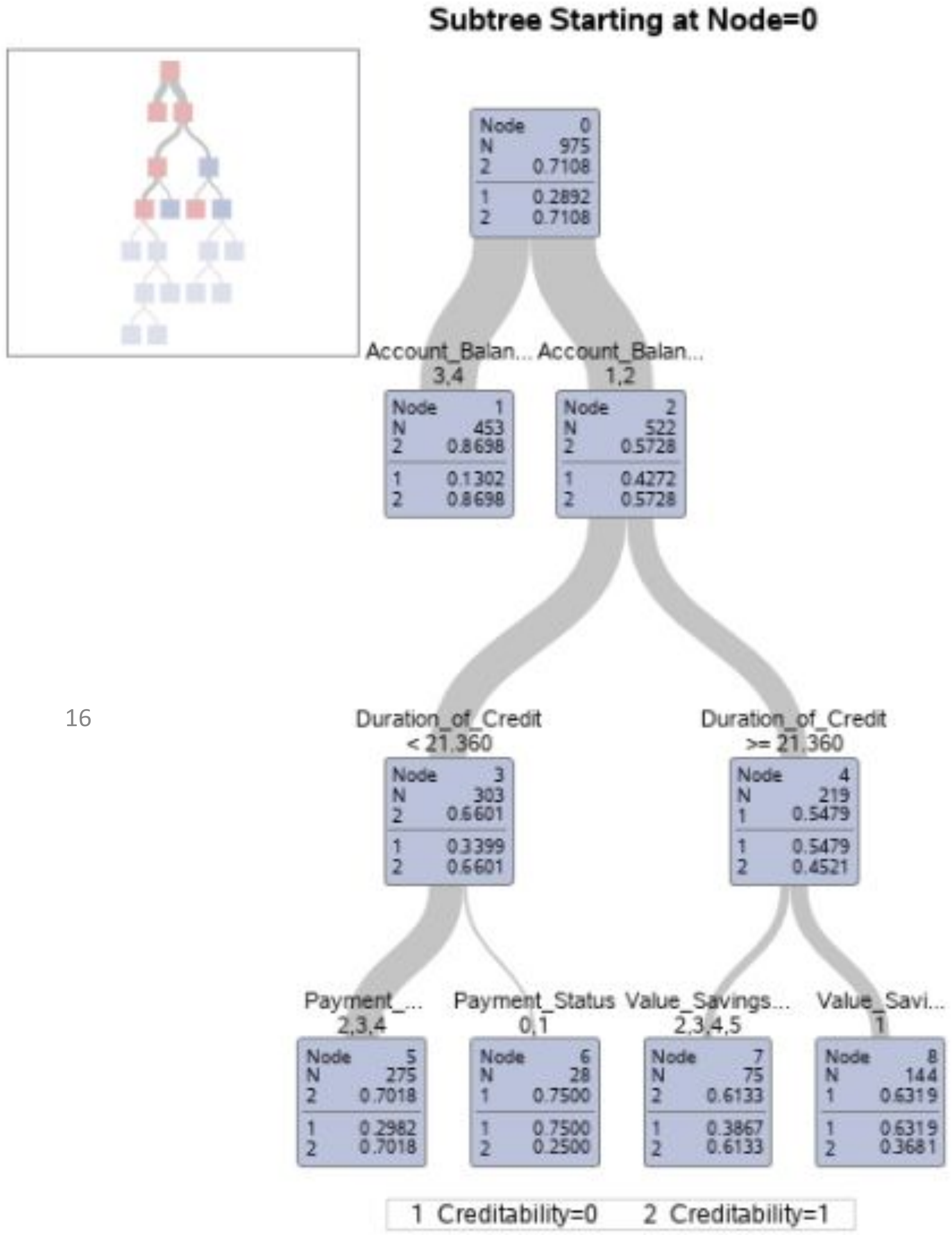
LLK Intelligence

Predictive Modelling - Decision Tree

Pre-processed data



Processed data



Predictive Modelling - Decision Tree

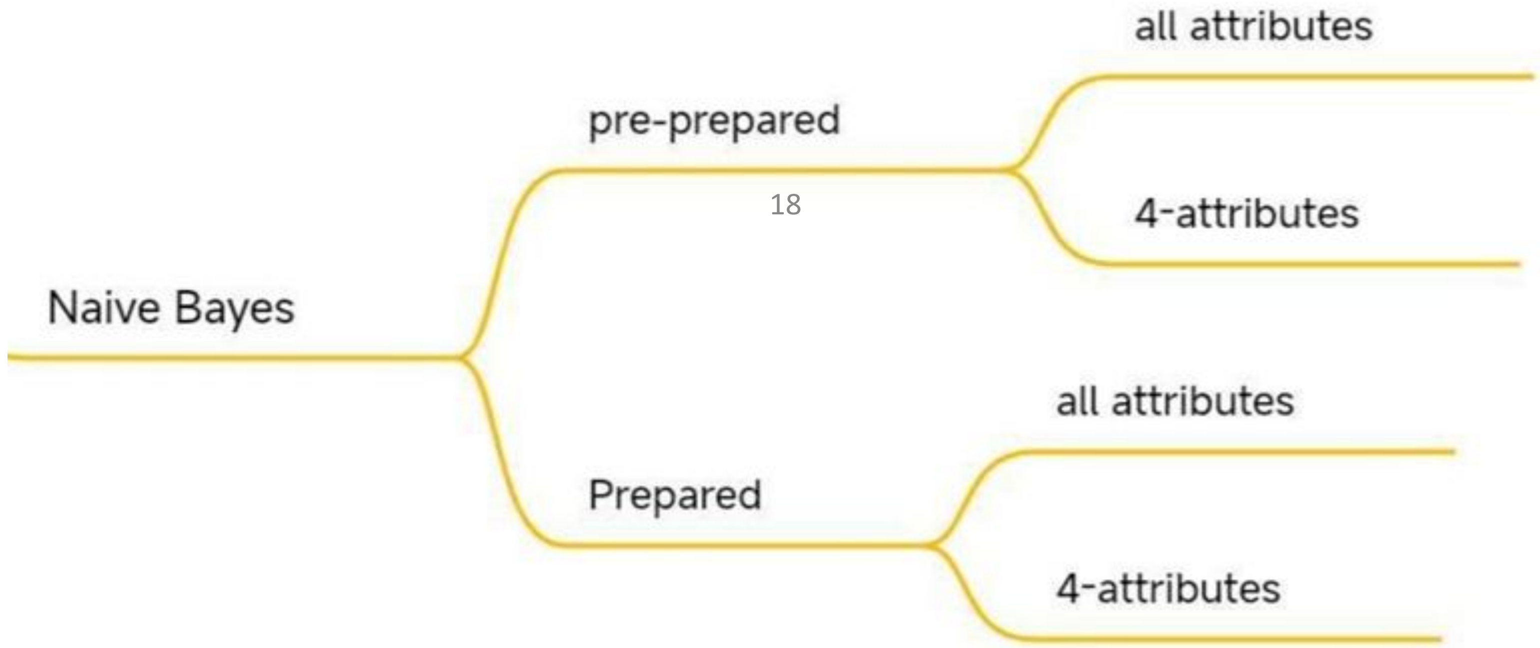
Based on ALL variables 0.1 trim					
Pre-prepared data			**Prepared data**		
Generated confusion matrix			Generated confusion matrix		
	0	1		0	1
0	38	68	0	35	49
1	16	178	1	24	185
Positive represents 1					
	+	-		+	-
+	178	16	+	185	24
-	68	38	-	49	35
Accuracy	0.72		Accuracy	0.750853	
Recall	0.917526		Recall	0.885167	
Precision	0.723577		Precision	0.790598	
F1 Score	0.809091		F1 Score	0.835214	

Based on ALL variables 0.25 trim					
Pre-prepared data			**Prepared data**		
Generated confusion matrix			Generated confusion matrix		
	0	1		0	1
0	40	66	0	49	35
1	39	155	1	39	170
Positive represents 1					
	+	-		+	-
+	155	39	+	170	39
-	66	40	-	35	49
Accuracy	0.65		Accuracy	0.74744	
Recall	0.798969		Recall	0.813397	
Precision	0.701357		Precision	0.829268	
F1 Score	0.746988		F1 Score	0.821256	

Based on 4 variables 0.25 trim					
Pre-prepared data			**Prepared data**		
Generated confusion matrix			Generated confusion matrix		
	0	1		0	1
0	52	54	0	40	44
1	20	174	1	31	178
Positive represents 1					
	+	-		+	-
+	174	20	+	178	31
-	54	52	-	44	40
Accuracy	0.753333		Accuracy	0.744027	
Recall	0.896907		Recall	0.851675	
Precision	0.763158		Precision	0.801802	
F1 Score	0.824645		F1 Score	0.825986	

Predictive Modeling

- Evaluation method: Naïve Bayes(Weka)
- The Split method:
- Train : Test = 70% : 30%



LLK Intelligence

Predictive Modelling - Naive Bayes

- Explain the Naïve Bayes model
- Bayes' theorem: $posterior = \frac{prior * likelihood}{evidence}$
- Naïve Bayes learnt the priors P_{class1}=0.7, P_{class0} =0.3. In our model:

```
Naive Bayes Classifier

                                        Class
Attribute                                   0       1
                                       (0.29)  (0.71)
=====================================================
Account Balance
  1                                      91.0    89.0
  2                                      64.0   105.0
  3                                      10.0    37.0
  4                                      37.0   257.0
  [total]                               202.0   488.0

Duration of Credit (month)
  mean                                23.8801 19.0948
  std. dev.                           12.7811 10.8691
  weight sum                              198     484
  precision                            1.8667  1.8667

Payment Status of Previous Credit
  0                                      16.0     9.0
  1                                      19.0    16.0
  2                                     113.0   245.0
  3                                      21.0    39.0
  4                                      34.0   180.0
  [total]                               203.0   489.0

Value Savings/Stocks
  1                                     149.0   271.0
  2                                      19.0    46.0
  3                                      10.0    36.0
  4                                       5.0    32.0
  5                                      20.0   104.0
  [total]                               203.0   489.0
```

Predictive Modelling - Naive Bayes

- Likelihood/Conditional Probabilities:
- Basically calculate each situation in each attributes given class 1 or class 0. using smoothing technology. +1 in molecular, +V in denominator, V means in the attribute, the number of all distinct situations in class1+ class0.

- Coose a class:
- Calculate each tuple as prior * all likelihood in two classes then compare.

Predictive Modelling - Naive Bayes

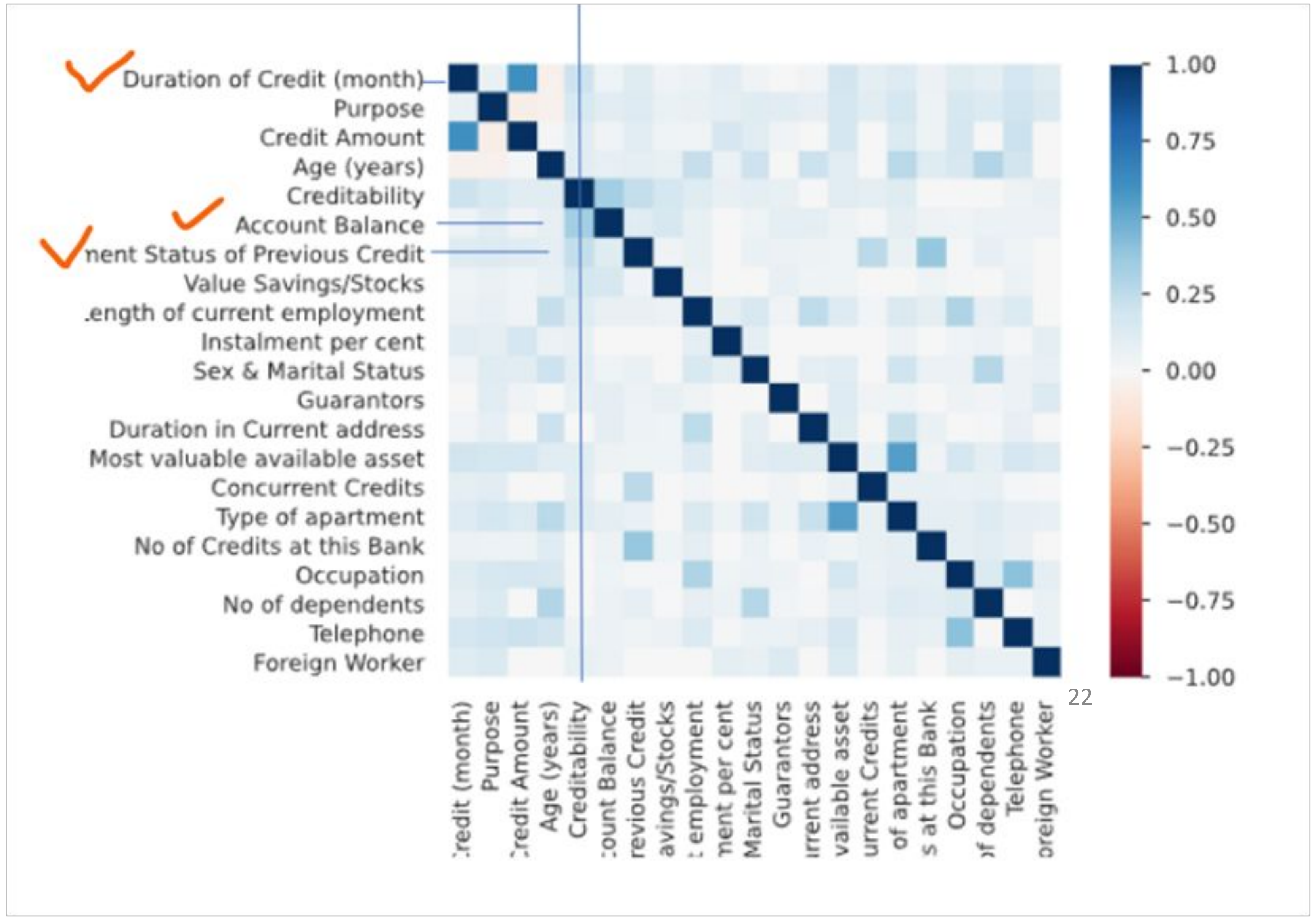
Based on ALL variables NB						
pre-prepared data				Prepared data		
	0	1			0	1
0	50	56		0	52	32
1	19	175		1	38	171
	1	0			1	0
1	175	19	11.19%	1	171	38
0	56	50		0	32	52
TP	0.902062	0.902062		TP	0.818182	0.818182
precision	0.757576	0.757576		precision	0.842365	0.842365
FP	0.528302	0.528302		FP	0.380952	0.380952
recall	0.902062	0.902062		recall	0.818182	0.818182
Accuracy	0.75	0.75		Accuracy	0.761092	0.761092

Based on 4 variables NB						
pre-prepared data				Prepared data		
	0	1			0	1
0	40	66		0	46	38
1	16	178		1	32	177
	1	0			1	0
1	178	16	12.85%	1	177	32
0	66	40		0	38	46
TP	0.917526	0.917526		TP	0.84689	0.84689
precision	0.729508	0.729508		precision	0.823256	0.823256
FP	0.622642	0.622642		FP	0.452381	0.452381
recall	0.917526	0.917526		recall	0.84689	0.84689
Accuracy	0.726667	0.726667		Accuracy	0.761092	0.761092

precision: Among the customers who are able to borrow in our predicted model, how many customers are actually able to borrow, and the rest are actually unable to lend
TP/Recall: Among all the customers who are actually able to borrow money, how many customers have we predicted(provide loan)

Exploratory Analysis

- Correlation Matrix(sk-learn)



Also shown the data after prepared, The most related attributes with the class attribute.

LLK Intelligence

Exploratory Analysis

- Linear regression(R studio)

```
lm(formula = Creditability ~ Account + Duration + Payment + Value,
    data = bank)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1053 -0.3861  0.1168  0.3047  0.7222 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.359636   0.049057   7.331 4.81e-13 ***
Account      0.100836   0.010861   9.285  < 2e-16 ***
Duration    -0.007214   0.001148  -6.284 4.99e-10 ***
Payment      0.064876   0.012440   5.215 2.24e-07 ***
Value        0.033884   0.008560   3.958 8.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4094 on 970 degrees of freedom
Multiple R-squared:  0.189,	Adjusted R-squared:  0.1856 
F-statistic: 56.51 on 4 and 970 DF,  p-value: < 2.2e-16
```

R-squared is much lower then 0.7, so they are not a good linear regression model.