



GERMAN CREDIT PREDICTION

Group 2



Name		Student ID
Quang Minh NGO	:	44284292
Tung Duong VUONG	:	44261535
Muhammad Umer	:	44135149
Aashish Agrawal	:	44009240

JUNE 7, 2019
MACQUARIE UNIVERSITY

Table of Contents

I. Background:	2
II. Situation:	3
III. Method	4
A. Overall Work flow	4
B. Data assessment	5
C. Data exploration	7
1. Correlation matrix	8
2. Frequency distribution checking for every variable	9
D. Predictive modelling for the whole dataset	10
E. Segmentation	12
F. Predictive modelling for each cluster	15
G. Compare and explain result	17
1. Comparing	18
2. Explanation:	19

Executive Summary

The report outlines the predictive models using demographic and socio-economic data to identify bad credit risk. The dataset available has no concerning errors and each attribute potentially provides nonoverlap information indicating that no data reduction is necessary. Thus, the whole data are used for analysis but we drop foreign worker variable for its uneven frequency distribution. We further proceed to analyze the data in two distinctive directions: using the whole data with clusters and without clusters. In the former approach, we found that the best predictive model is logistics regression by referring to cost gain matrix. In the latter approach, we clustered data by K-mean clustering method. As a result, we are able to point out three main clusters with different characteristics. Then, the dataset is divided in three parts in alignment with three clusters; and predictive analysis was conducted using each cluster data. The result just shows slightly better performance, however, we might run into overfitting problem and there are too few data point in one of cluster. In short, the best option is the logistic regression in the whole dataset.

I. Background:

One of the largest revenue generation in banking is the activity of lending to individuals. By doing so, our bank can have concrete money inflows from the interests earned from such loans. However, the biggest challenge in this business area is how to identify bad credit risk to avoid capital loss for our bank.

This report highlights the predictive analysis methods and outcomes based on the applicant's demographic and socio-economic profile to address the risk of loan approval to a customer.

II. Situation:

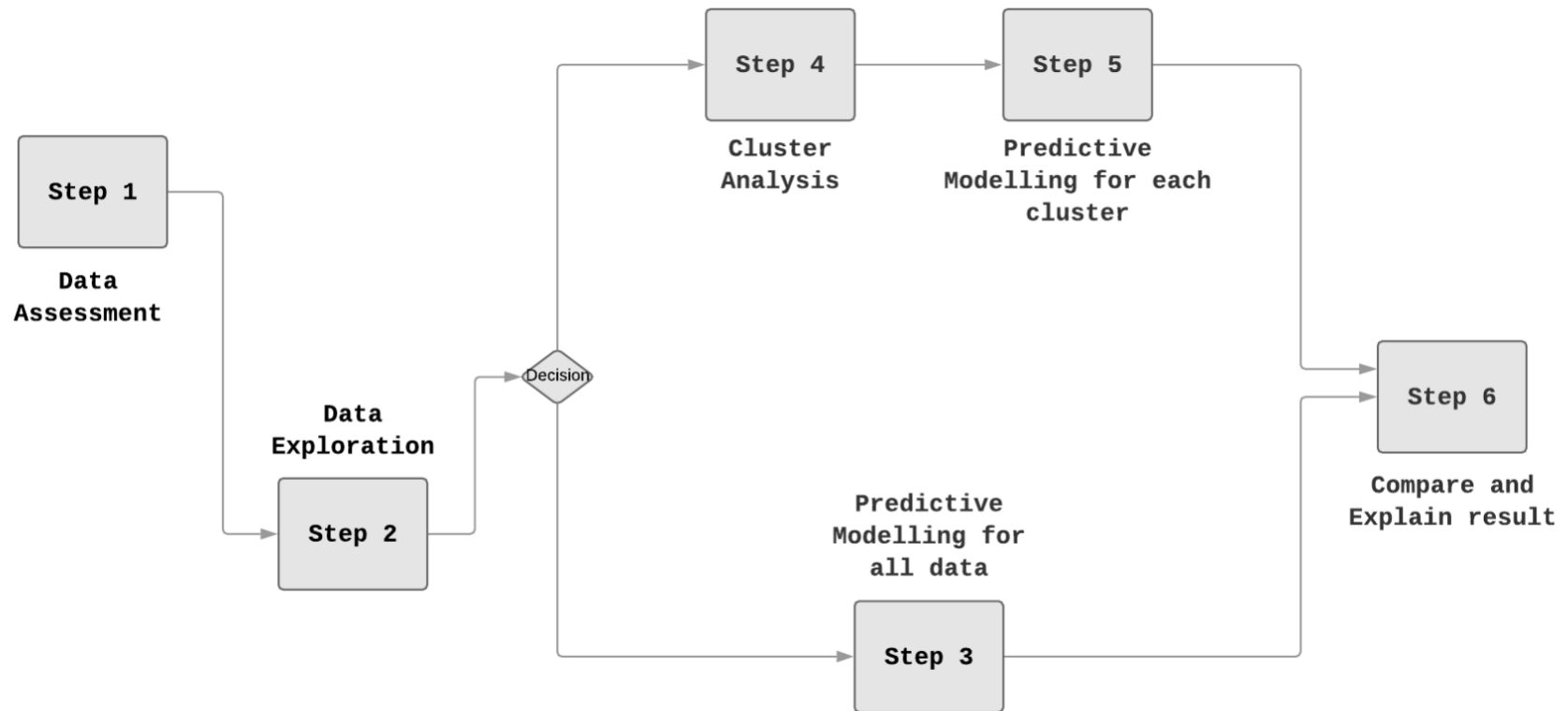
Currently, the dataset we collected consists of 1000 cases with 23 attributes which mainly are demography and financial standing of customers. Also noted, for each case, we know which one is good credit and which one is not.

Though, basing on available information we need to figure out:

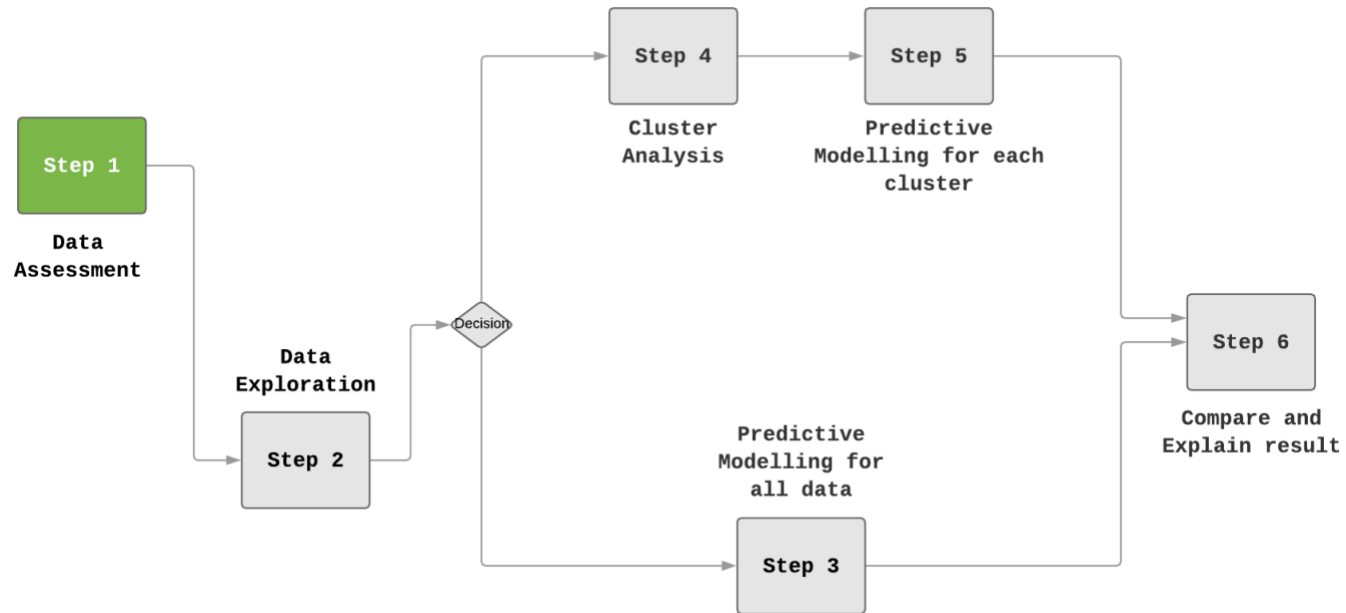
- A model with highest predictive ability of credit worthiness
- Tradeoff ratio to maximize business benefits
- What are market segments among our customer, will they make the prediction more accurate?

III. Method

A. Overall Work flow



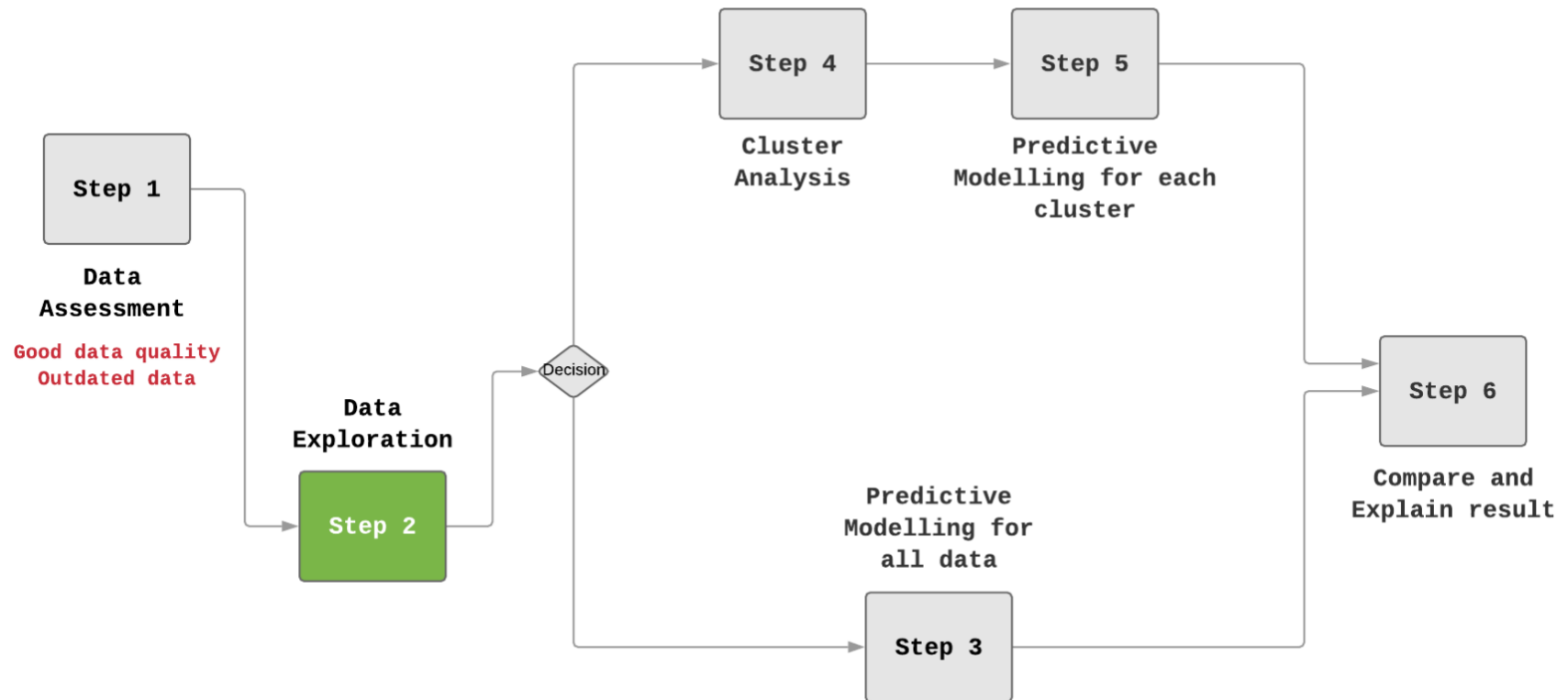
B. Data assessment



The data are assessed based on 6 dimensions namely, the completeness, uniqueness, timeliness, accuracy, consistency and validity.

Completeness	Uniqueness	Timeliness
The data are complete, no missing values are found	All cases have distinctive values, no duplication found	The data originates back in 1994, which are not up to date
Accuracy	Consistency	Validity
Unable to assess the accuracy of data, but let assume there are no wrong values in the dataset	Data are consistent throughout every row. (check by referring to the metadata dictionary)	0% of invalid data (compare to the metadata document)

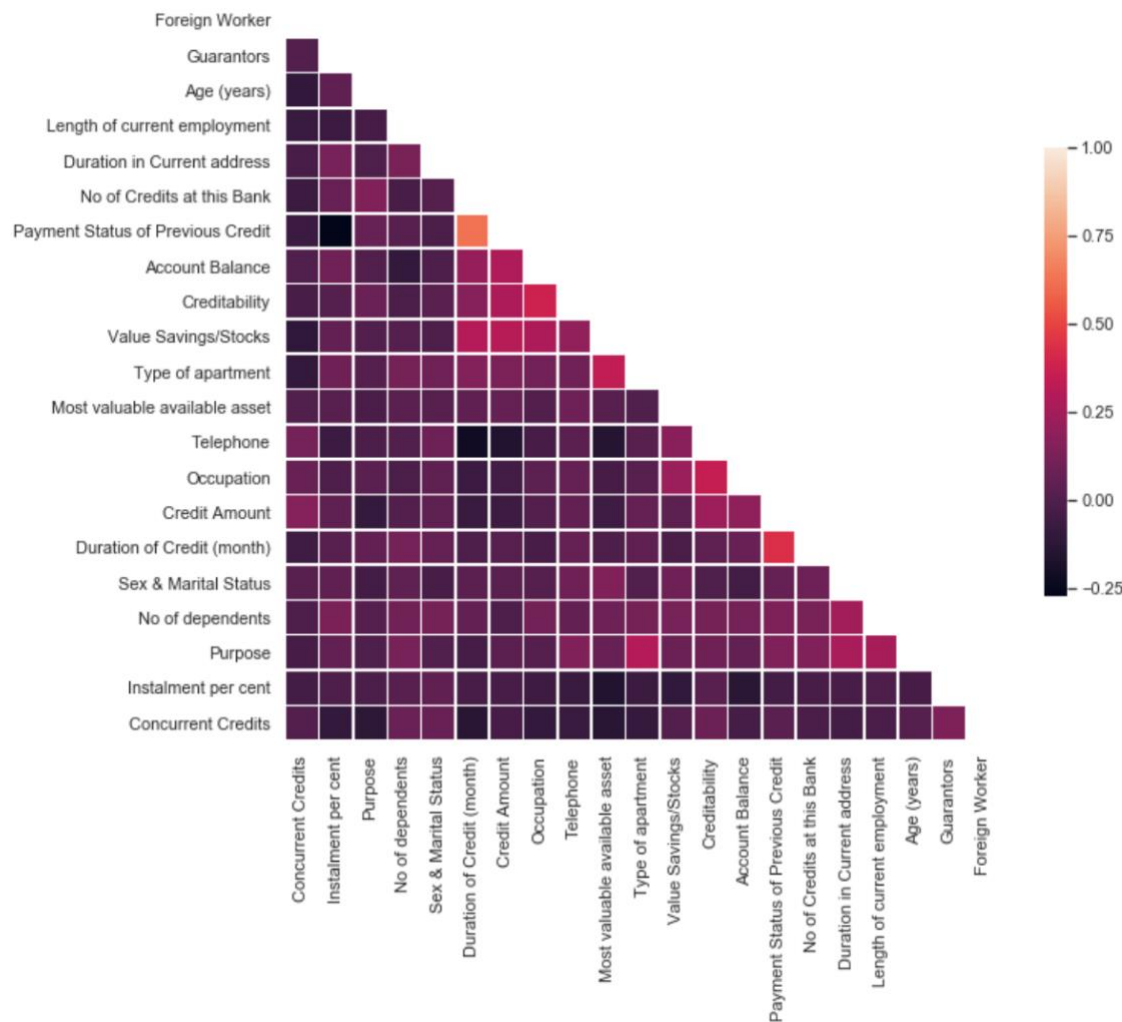
C. Data exploration



1. Correlation matrix

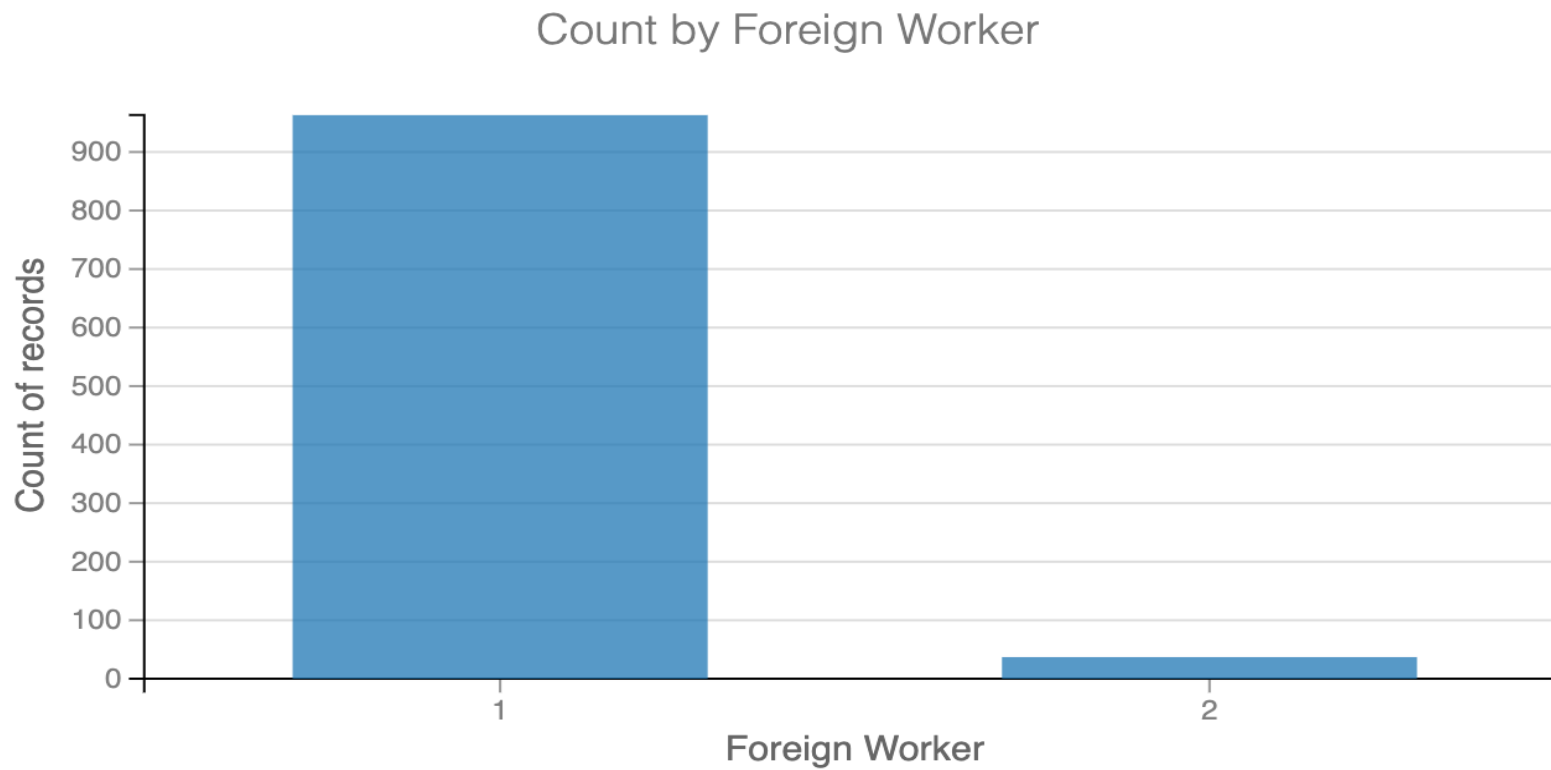
By examining the relationship between each variable, there is not many significant correlations between them except for payment status of previous credit and duration of credit. Thus, **data reduction will not be performed** as it will cause the loss of information.

Correlation matrix

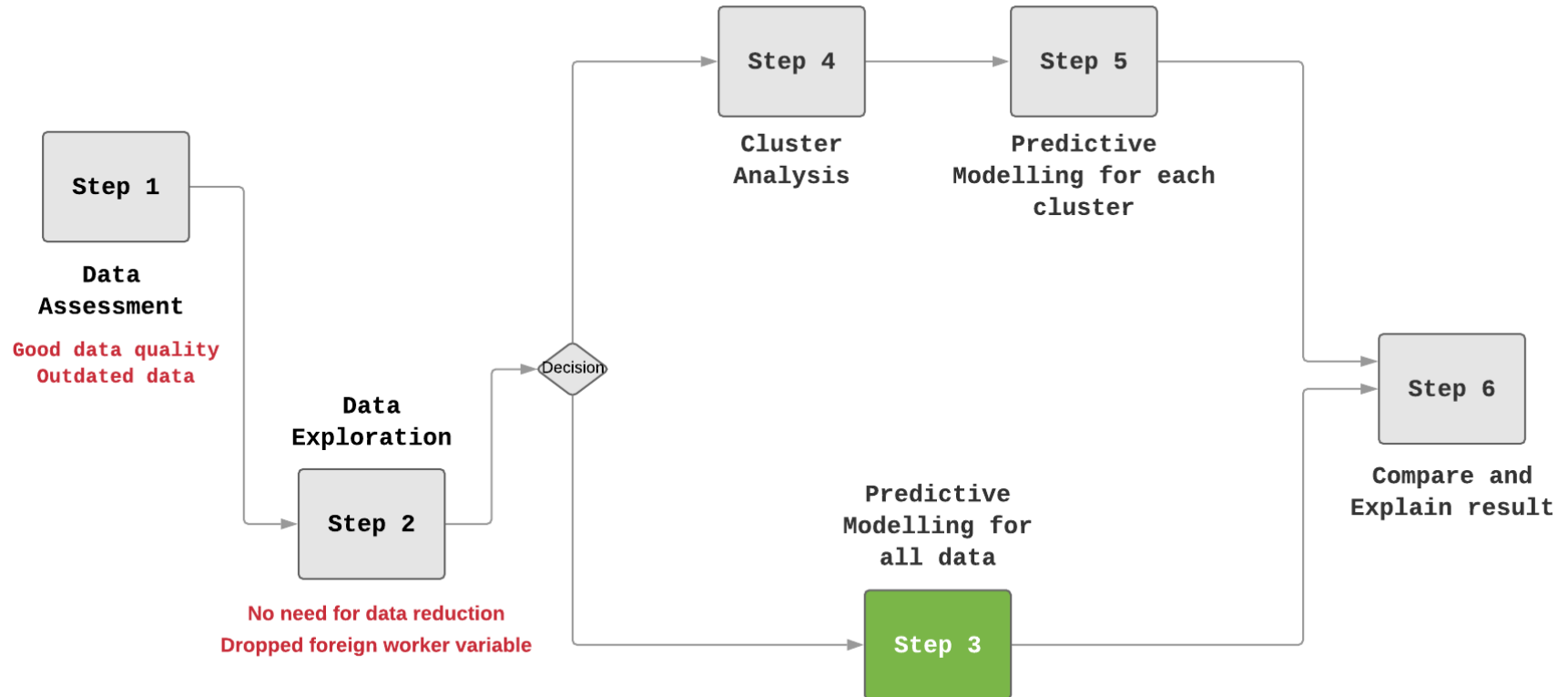


2. Frequency distribution checking for every variable

We check the number of records of each variable and found out there are too few customers who are not foreign workers, so **we would drop foreign workers variable.**



D. Predictive modelling for the whole dataset

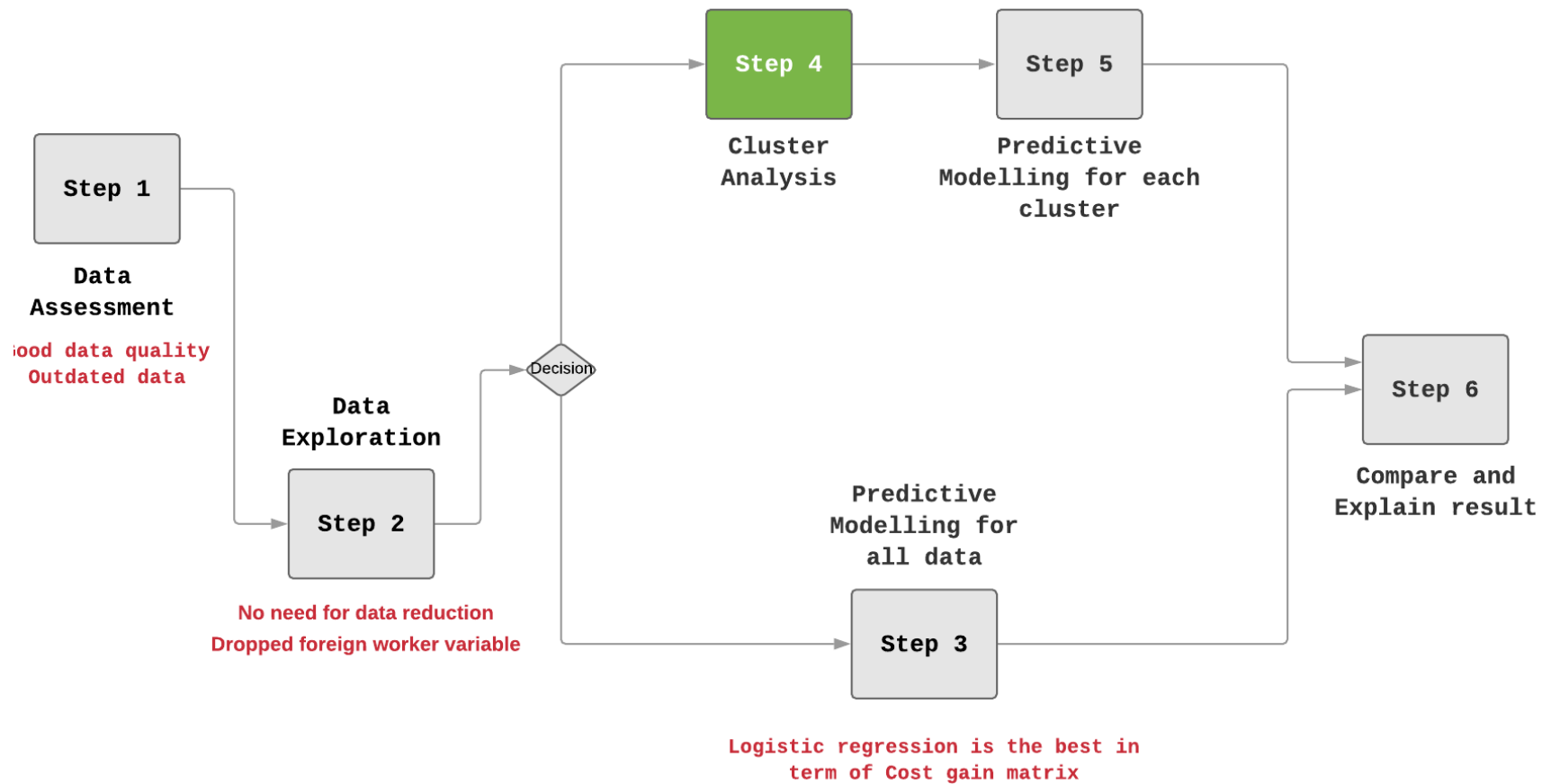


Training info
Full data set
Maximize Cost gain Matrix
Training records: 700
Testing records: 300

Methods	Cost Matrix
Logistic Regression	-136
Random forest	-144
SVM	-144
Artificial Neural Network	-147
Gradient Boosted Trees	-150
Extra trees	-159
XGBoost	-159
K Nearest Neighbors (k=5)	-189
Decision Tree	-195
SGD	-261

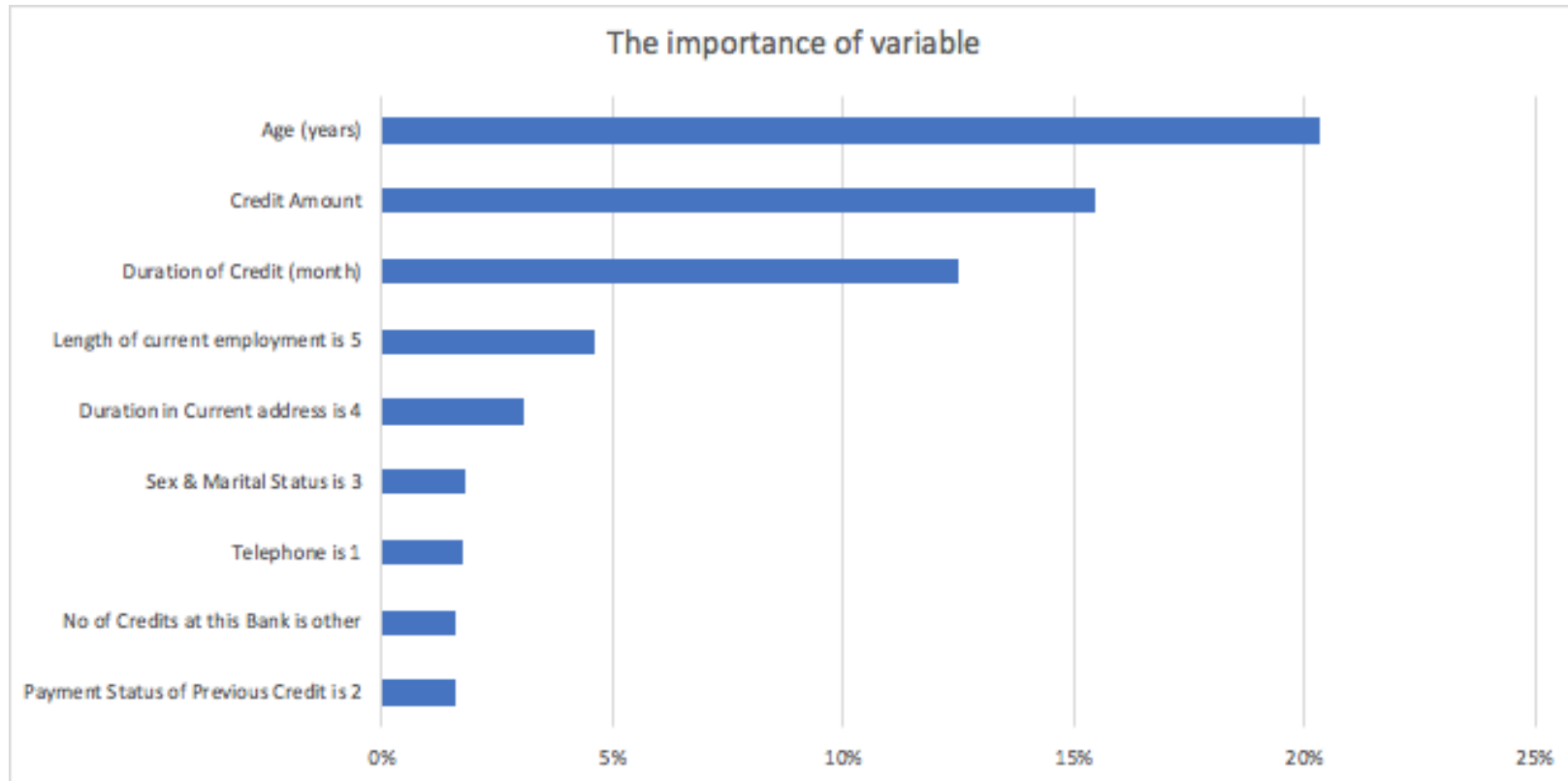
Logistic Regression works best in term of cost gain matrix

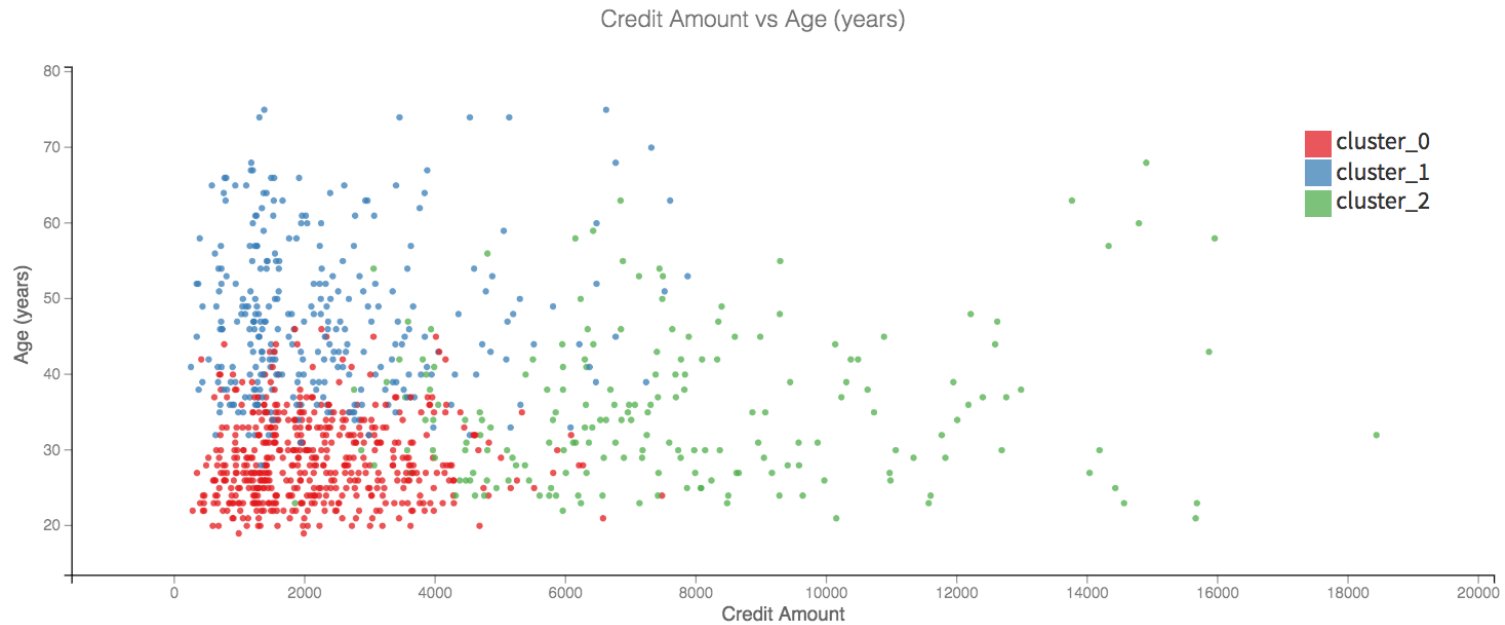
E. Segmentation



The purpose of segmentation is that we try to figure out the groups of people which have the same characteristics. By applying K-mean clustering method with 3 clusters, three distinctive clusters are found with different traits.

The most important variable in defining each cluster are the age, credit amount and duration of credit.



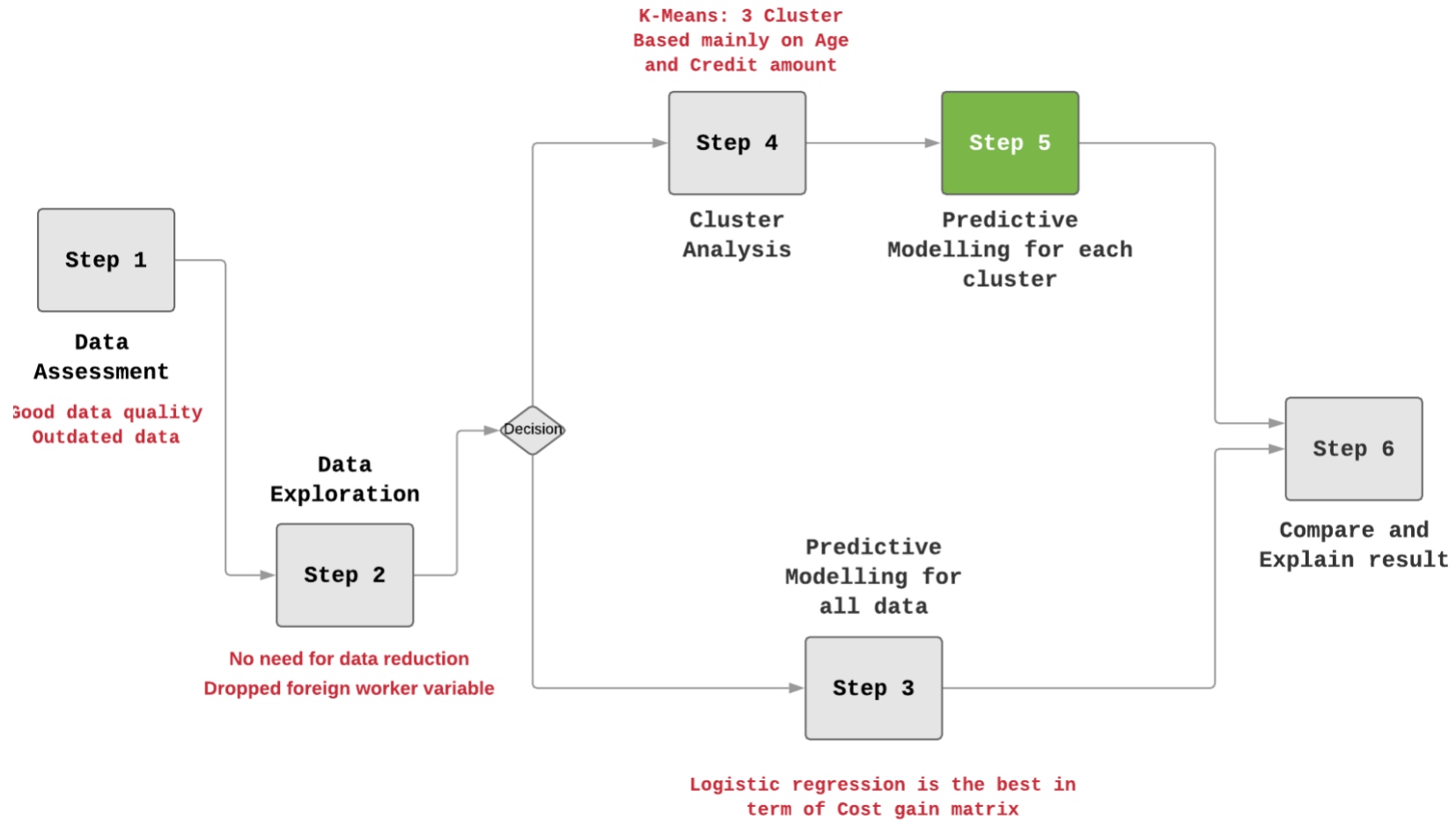


Scatter plot of three clusters based on age and credit amount

Cluster Profile

Cluster 0	Cluster 1	Cluster 2
They are young adults; most are under 40 years old. They have low credit amount and duration	They are senior customers. Same credit amount as cluster 0, but they have existing unpaid credit	Their age distributes widely, they have very high credit amount and duration of credit

F. Predictive modelling for each cluster



Training info
Maximize Cost gain Matrix

Cluster 0
Training records: 360
Testing records: 151

Method	Cost matrix
SVM	-67
Random forest	-69
Logistic Regression	-70
Extra trees	-71
Artificial Neural Network	-71
SGD	-72
XGBoost	-80
Gradient Boosted Trees	-89
K Nearest Neighbors (k=5)	-89
Decision Tree	-92

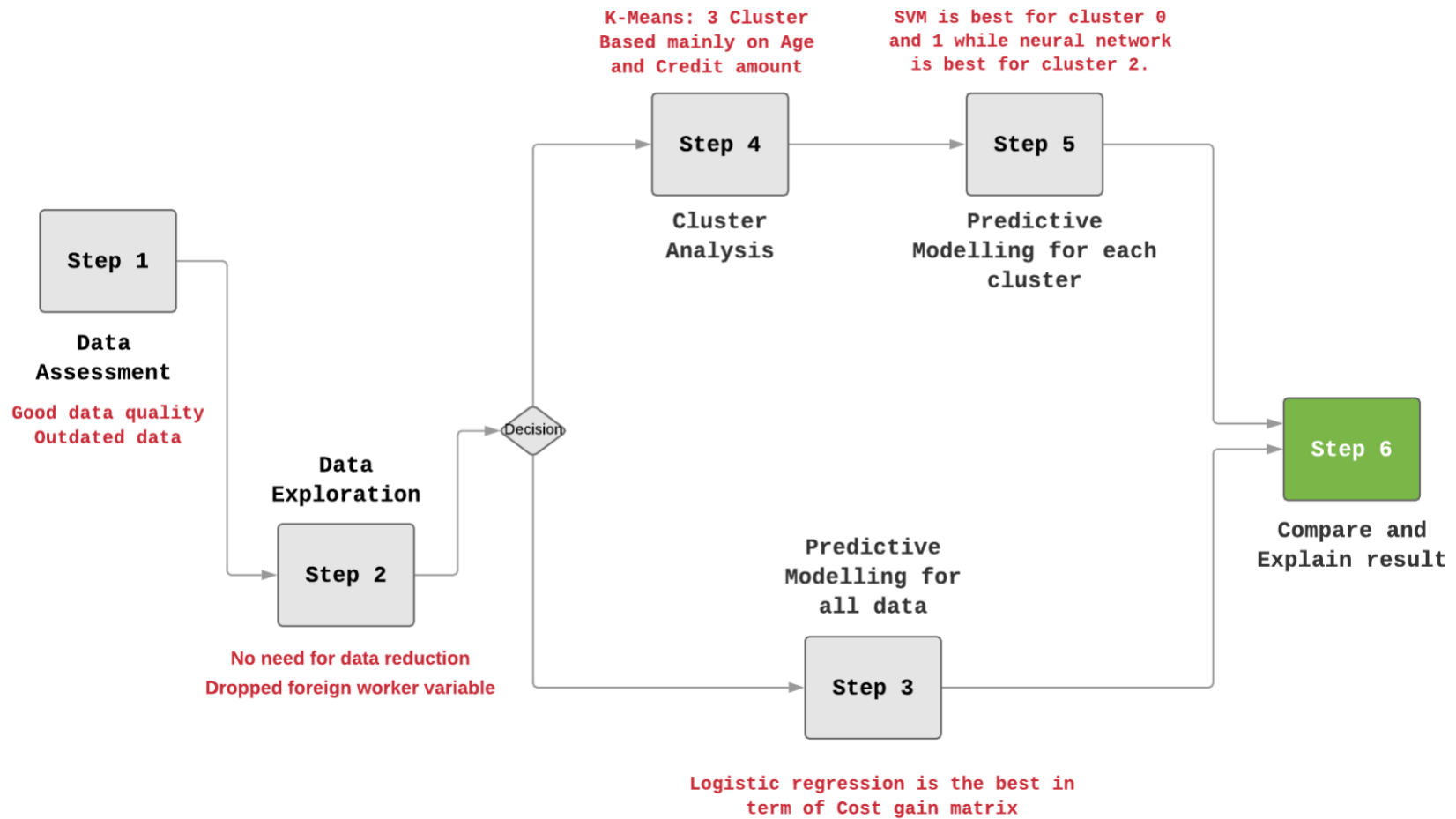
Cluster 1
Training records: 200
Testing records: 100

Method	Cost matrix
SVM	-45
Extra trees	-48
Random forest	-48
Artificial Neural Network	-50
Gradient Boosted Trees	-51
Logistic Regression	-51
K Nearest Neighbors (k=5)	-55
XGBoost	-56
Decision Tree	-66
SGD	-69

Cluster 2
Training records: 140
Testing records: 49

Method	Cost matrix
Artificial Neural Network	-16
XGBoost	-18
Logistic Regression	-18
SVM	-19
Random forest	-20
Extra trees	-21
Gradient Boosted Trees	-21
K Nearest Neighbors (k=5)	-22
Decision Tree	-40
SGD	-44


G. Compare and explain result



1. Comparing

Comparing in term of cost gain result

Without Clustering		With Clustering	
Best method	Cost gain	Best method	Combined Cost gain
Logistic Regression	-136	Mix (SVM + Artificial Neural Network)	-128



Total cost gain of 3 clusters with its best predictive analysis

- Even though using predictive modelling after clustering show better cost gain, the different between these 2 results is not much
- There are potential problems associate with clustering method like over fitting since we are doing 3 predictive modelling for 3 clusters, each of the cluster would have much lower number of records compare to the whole data set.
- There is one cluster with only 49 testing records and
→ **Therefore, it is not worth it clustering data first and then do predictive modelling.**
We want to just use logistic regression for the whole data set as it also easy to interpret as well

2. Explanation:

The following table show top 11 largest absolute coefficient of the logistic regression function

Variable	Coefficient
Account Balance is 4	0.73
Account Balance is 1	-0.62
Payment Status of Previous Credit is 4	0.49
Purpose is 1	0.49
Account Balance is 2	-0.39
Purpose is 3	0.36
Value Savings/Stocks is 5	0.36
Value Savings/Stocks is 1	-0.35
Purpose is 0	-0.35
Payment Status of Previous Credit is 1	-0.30
Purpose is 6	-0.24
Purpose is 9	-0.23

Looking at top 11 absolute value of coefficient, we could see following variables are affect the decision the most:

- **Account Balance**
- **Purpose**
- **Payment Status of Previous Credit**
- **Value Saving/Stocks**

IV. Conclusion:

Throughout the analysis, three customer segments are discovered among 1000 customer profile. Predictive models are then constructed within each of the segment. Though, the result indicates a little better performance, we do not have sufficient data point to reliably choose this method. Thus, for future research, we have to compile a larger dataset in order to proceed with this approach. However, at the moment, logistic regression is the best predictive modelling algorithm in term of cost gain matrix for the whole dataset