Default of Credit Card Analysis

Group 9

Yuying Chen, Syed Hassan Raza, Brittany Thum, Xinpei Zhao

Executive Summary

Project and Variable Overview

Predict loan defaults for 30,000 customers.

Exploratory Data Analysis

Exploration of the variables and their relationship with loan default

Age Segmentation

4 Age Segments

Project and Variable Overview

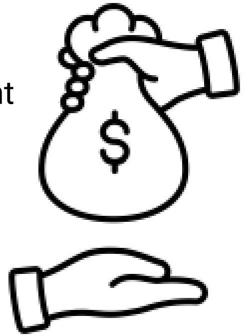
Data Background

Variable Overview

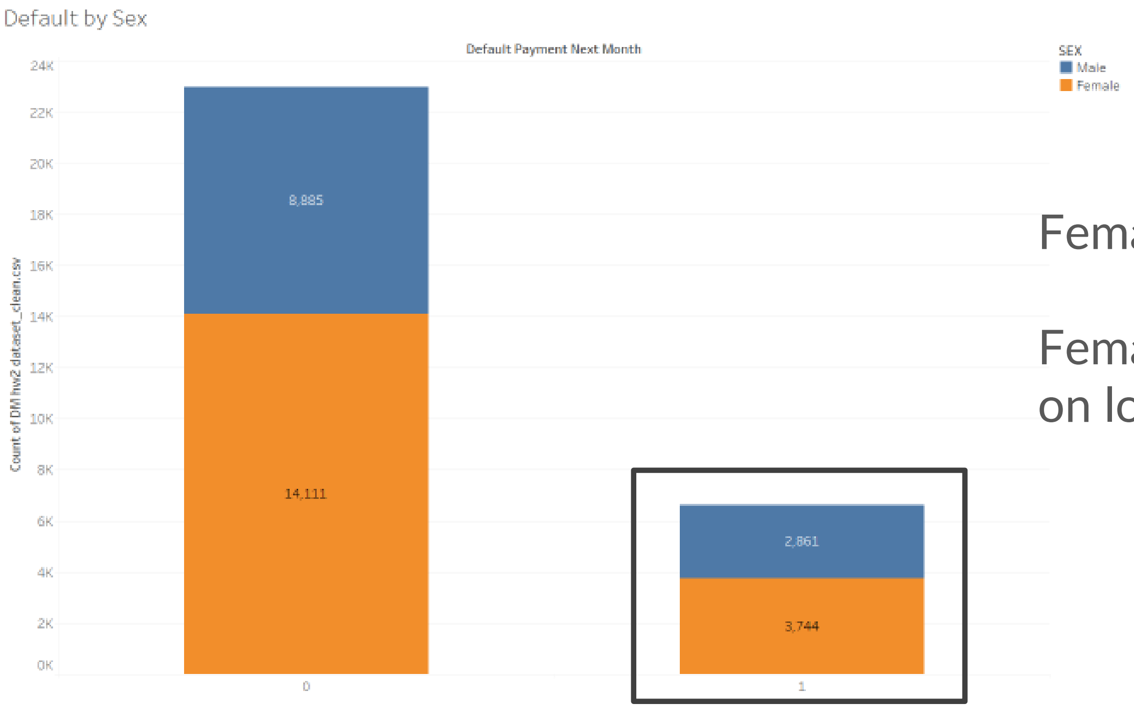
Data Structures: 30,000 Customers

25 Variables

LIMIT_BAL: Amount of the given credit (NT dollar)
SEX: Male and Female
EDUCATION: Graduation School, University, High School, and others
AGE: Age (year)
PAY_0 – PAY_6: History of past payment (Repayment status, monthly)
BILL_AMT1 – BILL_AMT6: Amount of bill statement (NT dollar, monthly)
PAY_AMT1 – PAY_AMT6: Amount of previous payment (NT dollar, monthly)



Which is the most common sex and which sex defaults the most?

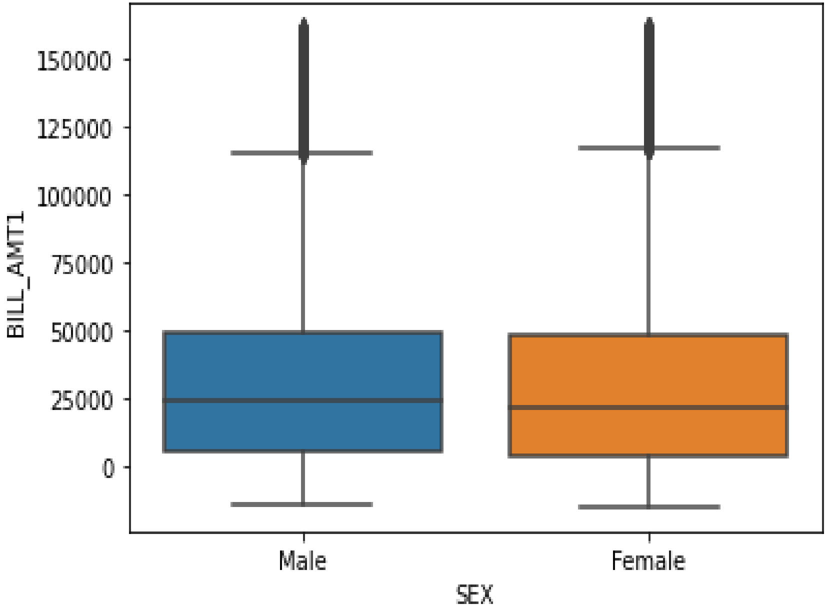


Female is the most common sex

Females are more likely to default on loans

How are Bill Amounts Distributed by Sex?

Bill Amounts are nearly equally distributed between Males and Females



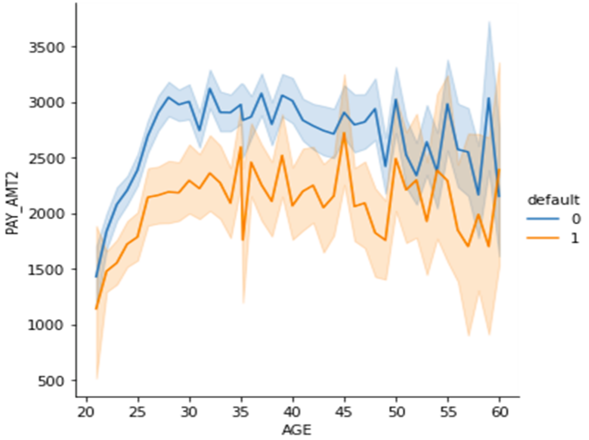
What is the relationship between Payment History and defaults?

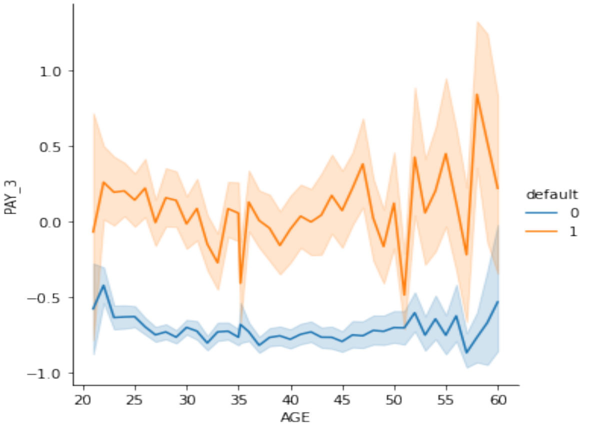
Delayed payment is directly proportional to default.

Higher the delay, the more likely to default.

If a client paid a higher amount last month, they are less likely to default the next month. The opposite is true, if they paid less last month, they are likely to default next month.

This relationship is true for all payment months.

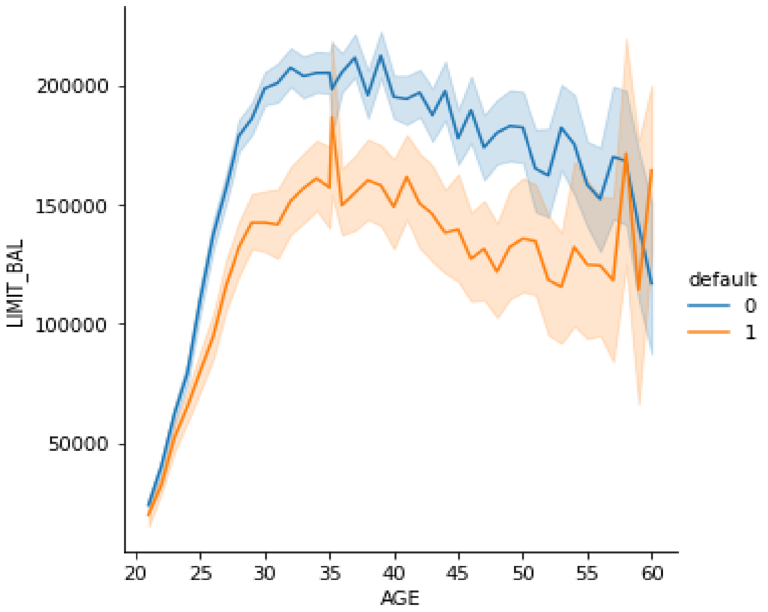




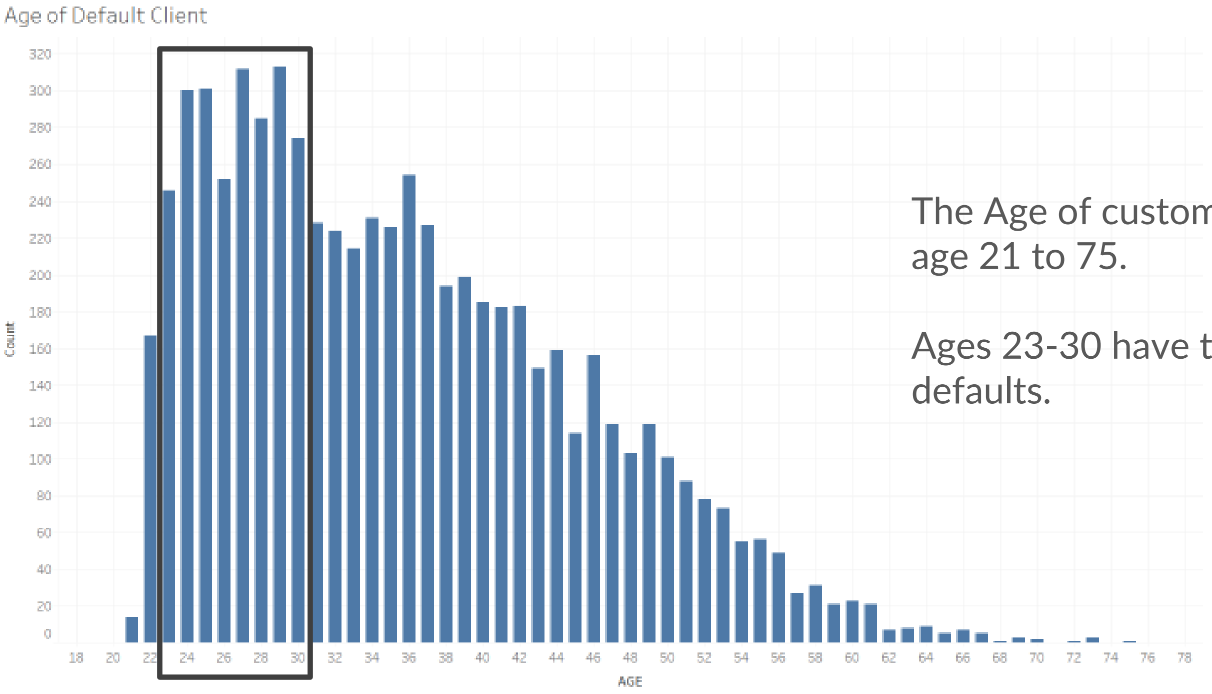
What is the relationship between Credit Limit and default?

Clients who have a higher credit limit are less likely to default.

These clients are already scrutinized by the bank for credit history and the ability to pay back.



Is there a relationship between age and default?



The Age of customers is distributed between age 21 to 75.

Ages 23-30 have the highest count of defaults.

What are the 4 Age Segments?

Predictors	Young Adults	Older Millennials	Middle Age Adults	Older Adults + Retirement
Age	21-27	28-34	35-44	45-75
Education	University	University	University	University
Relationship	Single	Single	Married	Married
Sex	Female	Female	Female	Female

Summary

Variable relationship with default	4 Age Segments
Females are more likely to default	The clustering resulted in 4 segments
There is a strong relationship between payments amounts and default	The bank can use the segments to aid in determining their loan process

Group 9

Appendix: Answers to Questions

Q1.1 How many customers are in the sample?

There are 30000 customers in the sample.

Row ID	Dimen...
Number Rows	30000
Number Col...	25

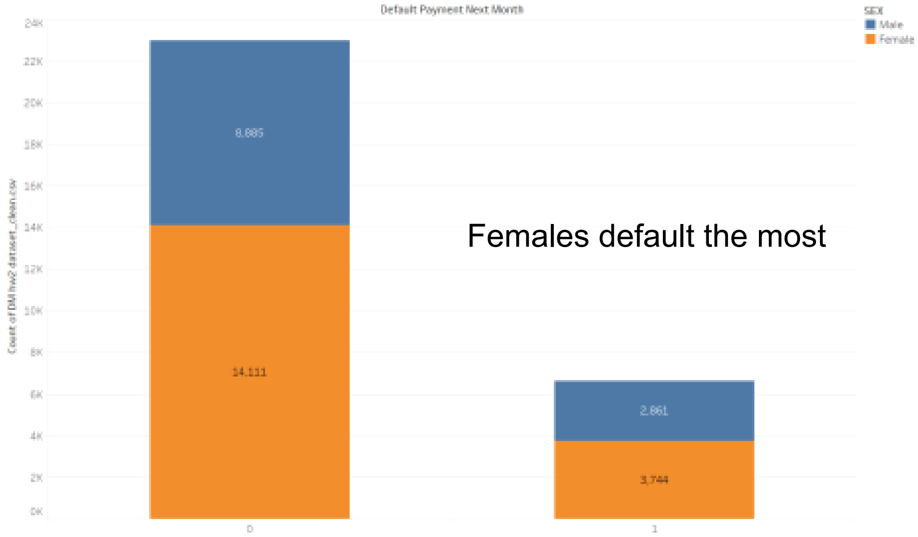
Q1.2 What is the most common sex in the sample?

Female is the most common sex in the sample

Row ID	count
Female	18112
Male	11888

Q1.3 Which sex has the most defaults?

Default by Sex



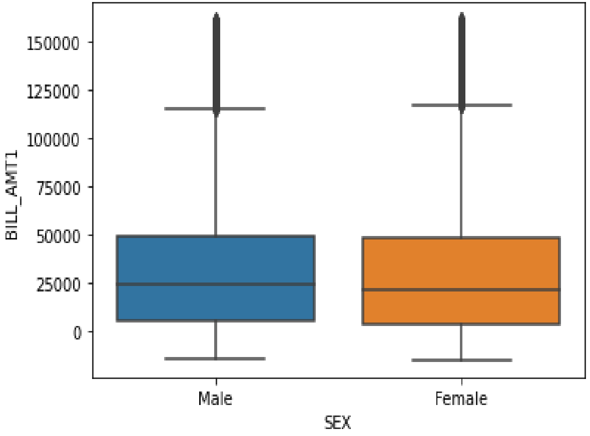
Females default the most

Q1.4 How many distinct values does marriage take on?

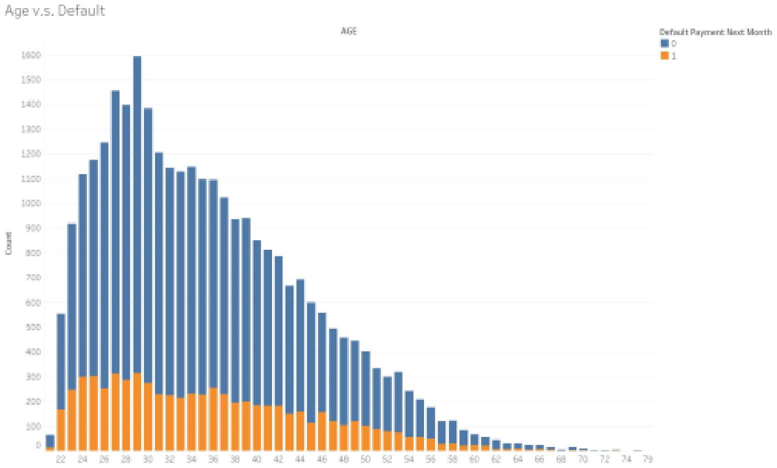
Row ID	Unique count*(MARRIAGE)
Row0	4

Marriage takes on 4 distinct values

Q2.1 How is BILL_AMT1 distributed by sex?



Q2.2 Does there appear to be any relationship between default and AGE?

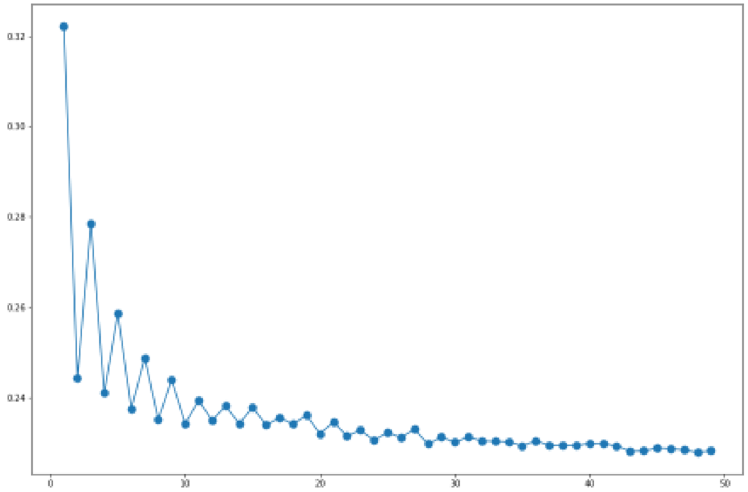


There seems to be no significant relationship between default and AGE.

Q3.1 Build a model of default using kNN. Randomly partition the data into a training set (70%) and a validation set (30%). What value of k did you decide to use and why?

k = 31 is our optimal value

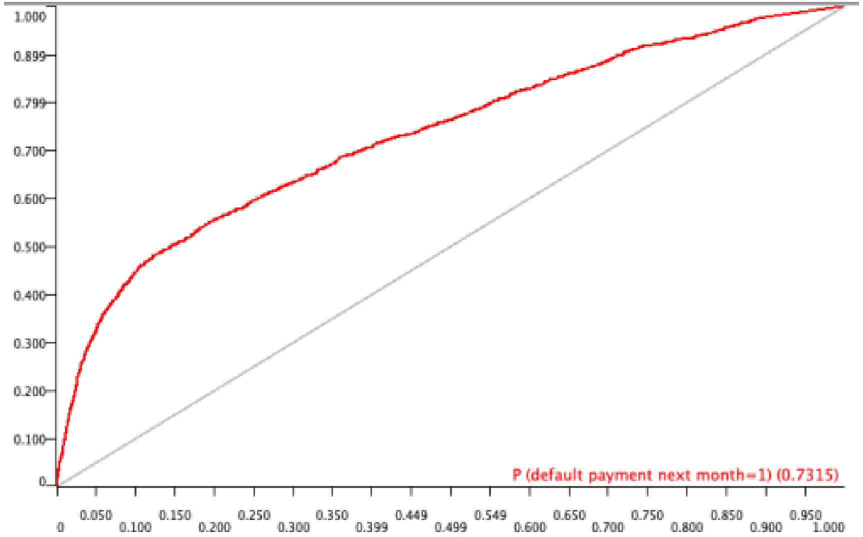
We calculated error rate at different k values and plotted the line graph of error rate against k value. At k =30 we are seeing the graph getting smoother at k ~= 30. And In order to avoid the tie, we are using odd value k = 31



Q3.2 Score the validation data (predict) using the model. Produce a confusion table and an ROC for the scored validation data.

Confusion Matrix

default pay...	1	0
1	688	1265
0	400	6528



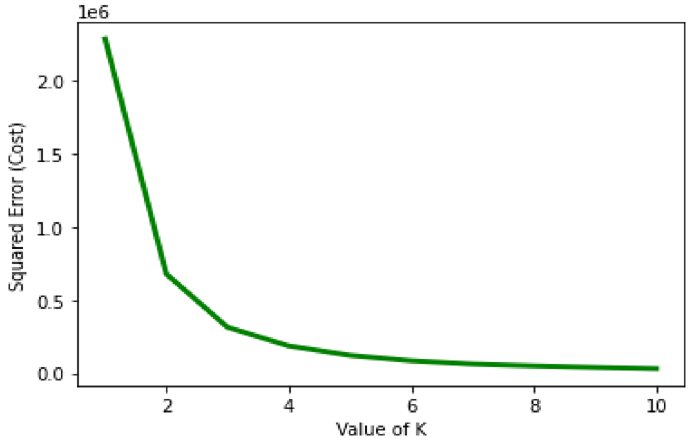
R score = 0.73

Q3.3 From the confusion table calculate the following metrics: accuracy, misclassification rate, true positive rate, false positive rate, specificity, precision, and prevalence.

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen...
1	84	113	6772	1912	0.042	0.426	0.042	0.984	0.077	?	?
0	6772	1912	84	113	0.984	0.78	0.984	0.042	0.87	?	?
Overall	?	?	?	?	?	?	?	?	?	0.772	0.038

Q3.4 Use k-means clustering to segment the customers on AGE. What value of k did you decide to use and why?

K = 4 is optimal value of k for our k-means clustering to segment the customers on AGE.
We calculated squared error or cost function and found that our error curve smooths after k =4



Q3.5 Build a model of default using kNN for each segment. Randomly partition the data into a training set (70%) and a validation set (30%) for each segment. What value of k did you decide to use and why?

Putting low value of can add lot of noise and that noise can influence our predictions. At the same time, putting very high value of k are computationally expensive. We received same result for values greater than 70 so, 55 was our choice.

Cluster_0: k = 55
Cluster_1: k = 65
Cluster_2: k = 65
Cluster_3: k = 73

Q3.6 Score the validation data (predict) using the models. Produce a confusion table for the scored validation data for each segment. How do they compare?

Cluster_0

default...	1	0
1	12	499
0	10	1437

Cluster_1

Row ID	1	0
1	5	546
0	9	2138

Cluster_2

Row ID	1	0
1	21	562
0	31	2055

Cluster_3

Row ID	1	0
1	15	376
0	15	1150

Q3.7 From the confusion tables for each segment calculate the following metrics: accuracy, misclassification rate, true positive rate, false positive rate, specificity, precision, and prevalence. How do they compare?

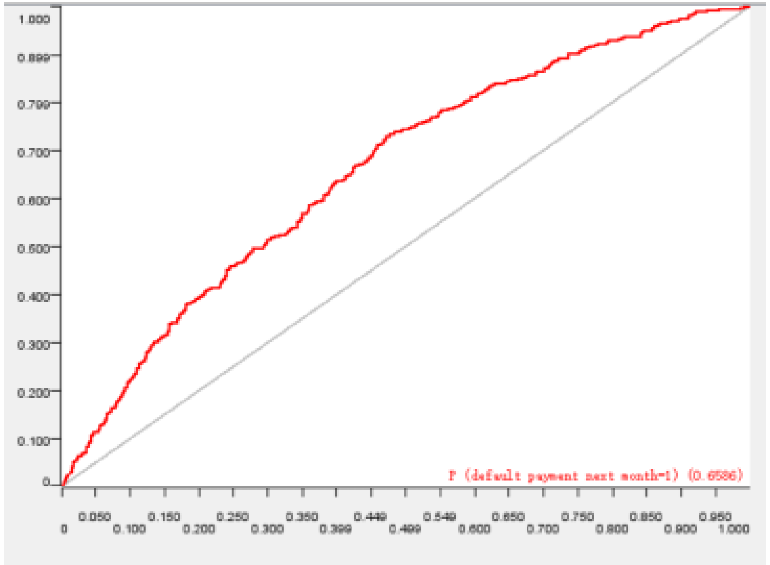
Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
1	12	10	1437	499	0. 023	0. 545	0. 023	0. 993	0. 045	?	?
0	1437	499	12	10	0. 993	0. 742	0. 993	0. 023	0. 85	?	?
Overall	?	?	?	?	?	?	?	?	?	0. 74	0. 024

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
1	5	9	2138	546	0. 009	0. 357	0. 009	0. 996	0. 018	?	?
0	2138	546	5	9	0. 996	0. 797	0. 996	0. 009	0. 885	?	?
Overall	?	?	?	?	?	?	?	?	?	0. 794	0. 008

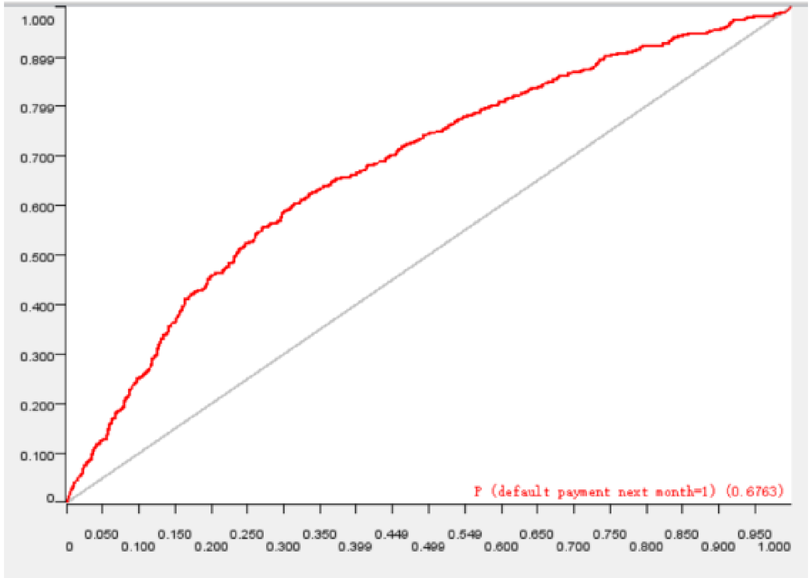
Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
1	21	31	2055	562	0. 036	0. 404	0. 036	0. 985	0. 066	?	?
0	2055	562	21	31	0. 985	0. 785	0. 985	0. 036	0. 874	?	?
Overall	?	?	?	?	?	?	?	?	?	0. 778	0. 031

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
1	15	15	1150	376	0. 038	0. 5	0. 038	0. 987	0. 071	?	?
0	1150	376	15	15	0. 987	0. 754	0. 987	0. 038	0. 855	?	?
Overall	?	?	?	?	?	?	?	?	?	0. 749	0. 037

Q3.8 Produce an ROC curve for each AGE segment and report the AUCs.

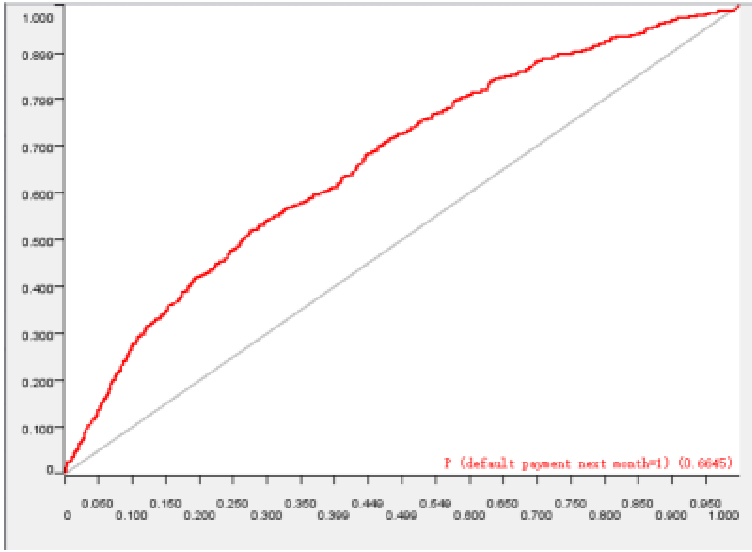


21-27

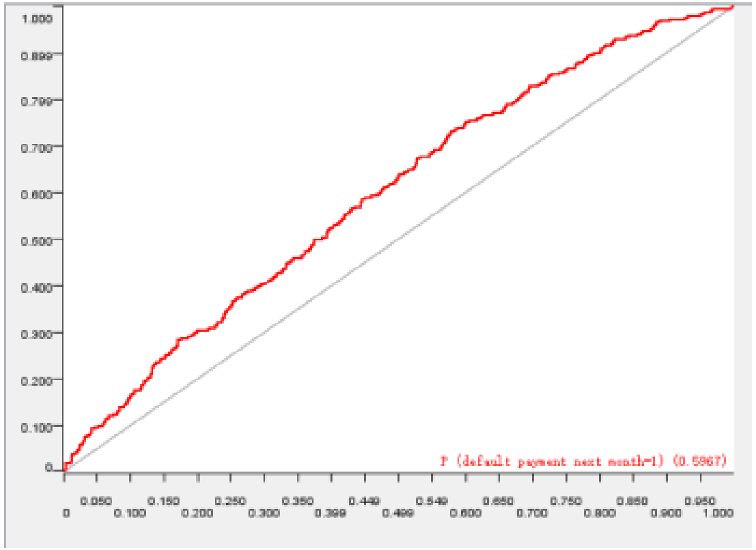


28-34

(cont.)Q3.8 Produce an ROC curve for each AGE segment and report the AUCs.



35-44



45-75

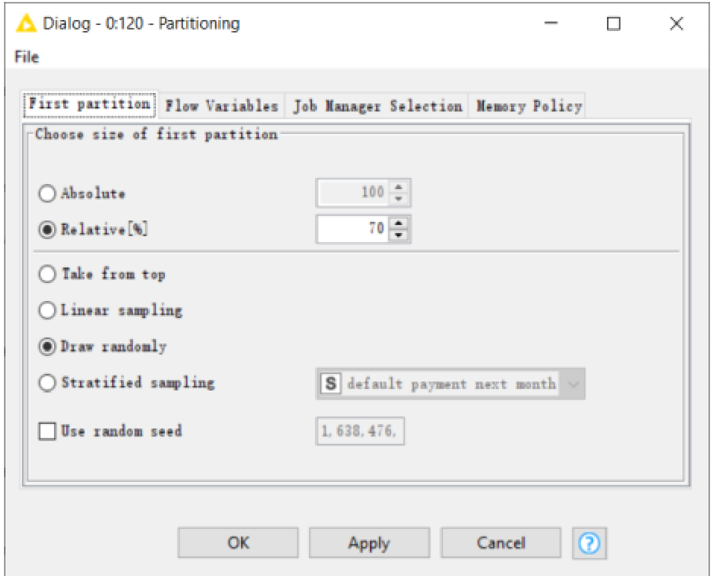
Q3.9 Do any of the models built on the AGE segments have a better classification performance than the non-segmented population model? How much better or worse?

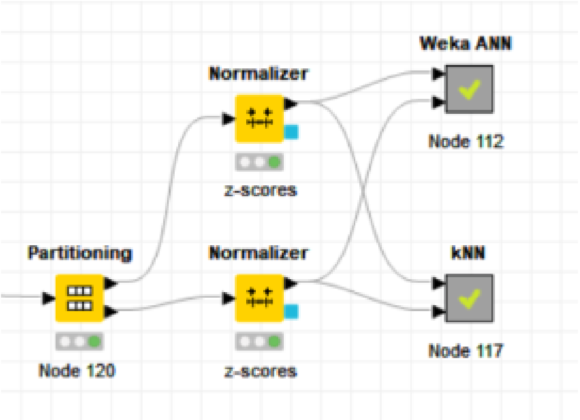
Non-Segmented Model Accuracy:
KNN : 77.2%
NN : 81.8%

Age Clustered KNNs:
74% , 79.4 , 77.8 , 74.9   (or Mean Accuracy = 76.52%)

Non-Segmented Models have slightly better classification accuracy than Model Built on AGE Segments

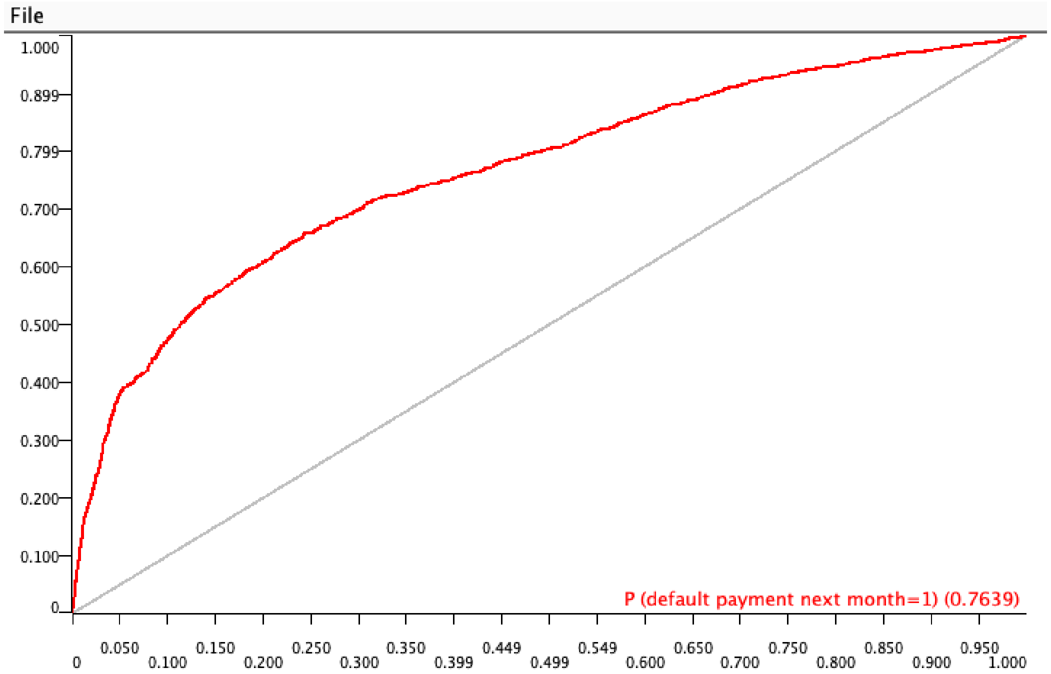
Q4.1 Build a model of default using ANN. Randomly partition the data into a training set (70%) and a validation set (30%).





Q4.2 Score the validation data (predict) using the model. Produce a confusion table and an ROC for the scored validation data.

Row ID	1	0
1	766	1272
0	340	6503



Q4.3 From the confusion table calculate the following metrics: accuracy, misclassification rate, true positive rate, false positive rate, specificity, precision, and prevalence.

Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensiti...	Specifi...	F-measure	Accuracy	Cohen's ...
1	766	340	6503	1272	0.376	0.693	0.376	0.95	0.487	?	?
0	6503	1272	766	340	0.95	0.836	0.95	0.376	0.89	?	?
Overall	?	?	?	?	?	?	?	?	?	0.818	0.389

Q5.1 Of the three models, which do you prefer to use and why?

Based on our ROC score and Accuracy Score results, we prefer Neural Networks for our results.

There are certain advantages Neural Network over KNN:

- Neural Networks have better ability to learn and model non-linear and complex relationships
- There are less restrictions on the input of Neural Networks
- Results of KNN depend on optimal value of K value that we need to find based on our model This K value may not suit if our model is tested on external data, thus overfitting the model

Accuracy Score of NN: 81.8 vs Accuracy Score of KNN (taken the best model) : 77.2%