LOAN DEFAULT PREDICTION

Business Question: Can we predict potential loan defaulters from a bank's customer data?

Yes, we can!

EXECUTIVE SUMMARY

- Potential loan defaulters can be accurately predicted from a bank's customer dataset
- ML algorithms that we implemented:
  - K-Nearest Neighbor
  - K-Means Clustering
  - Neural Networks
- **Best algorithm for prediction:**
  - Neural Network Algorithm

PREDICTOR VARIABLES

Pay 1 – Pay 6 :

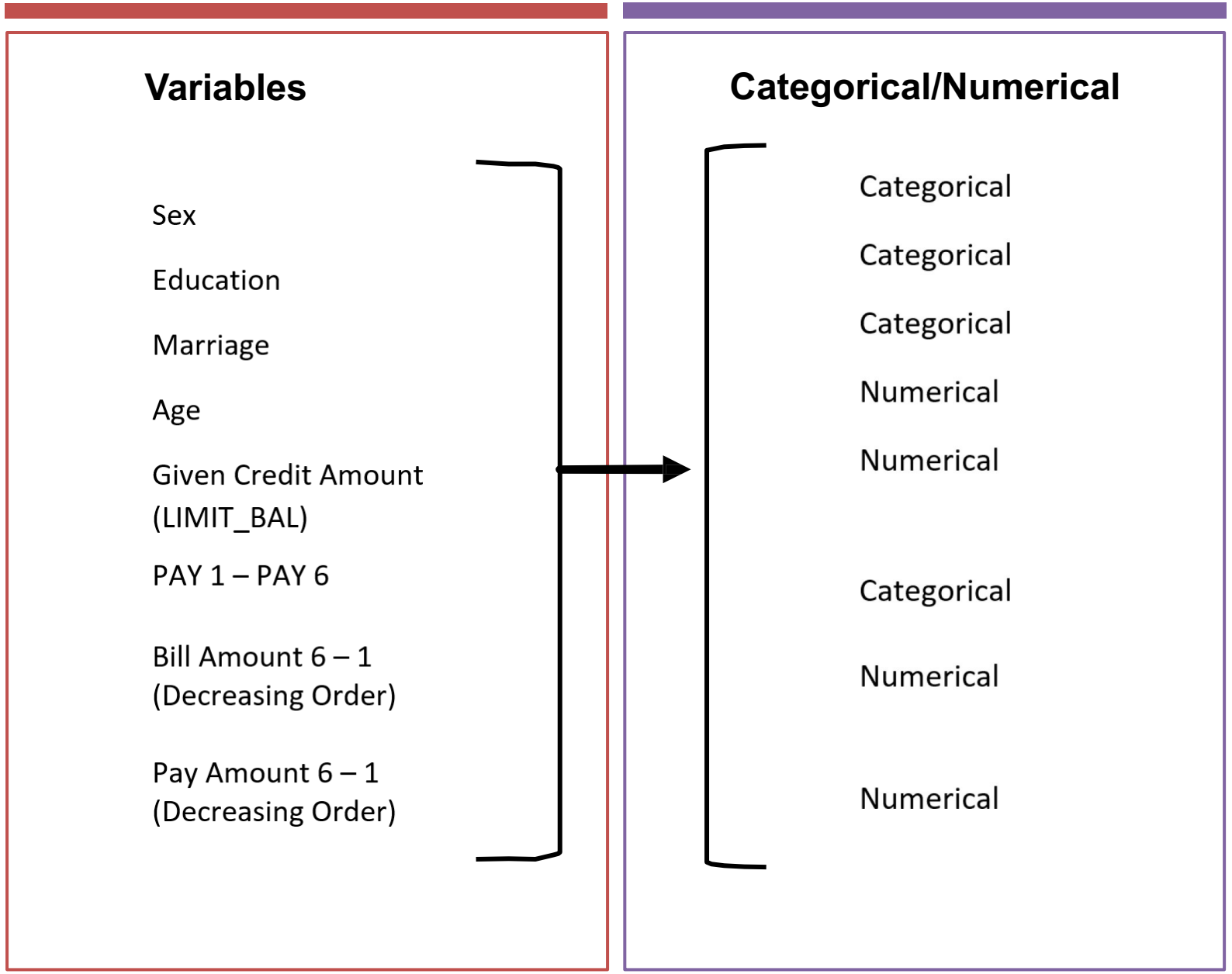
Repayment status from April to September 2005

Bill Amount 6 – Bill Amount 1 :

Bill statement amount from April to September 2005

Pay Amount 6 – Pay Amount 1 :

Previous payment amount from April to September 2005

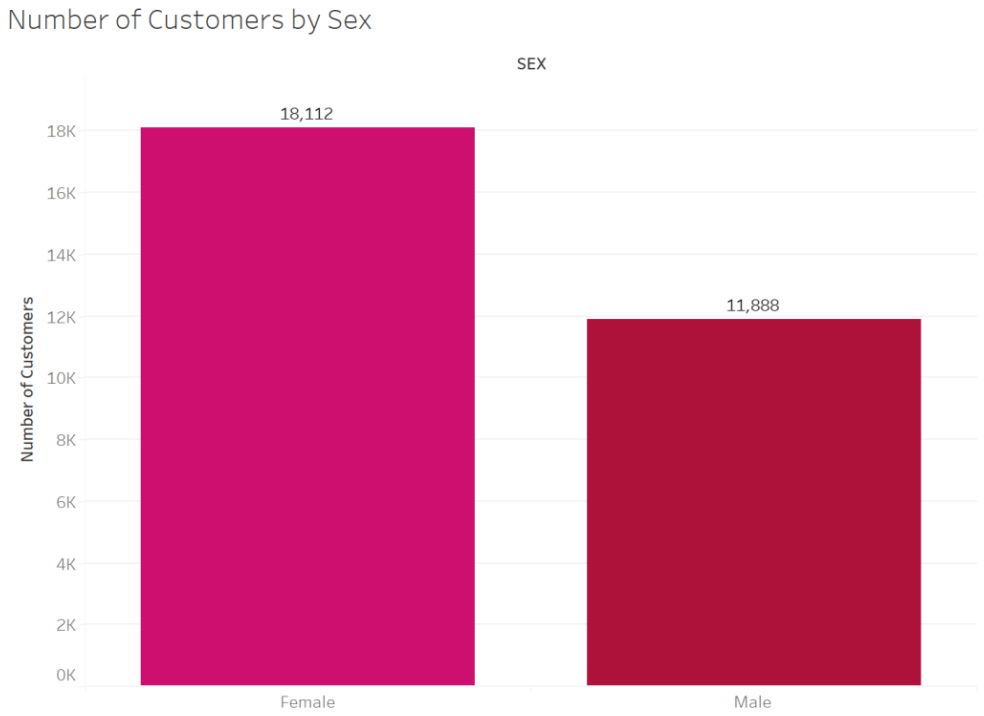


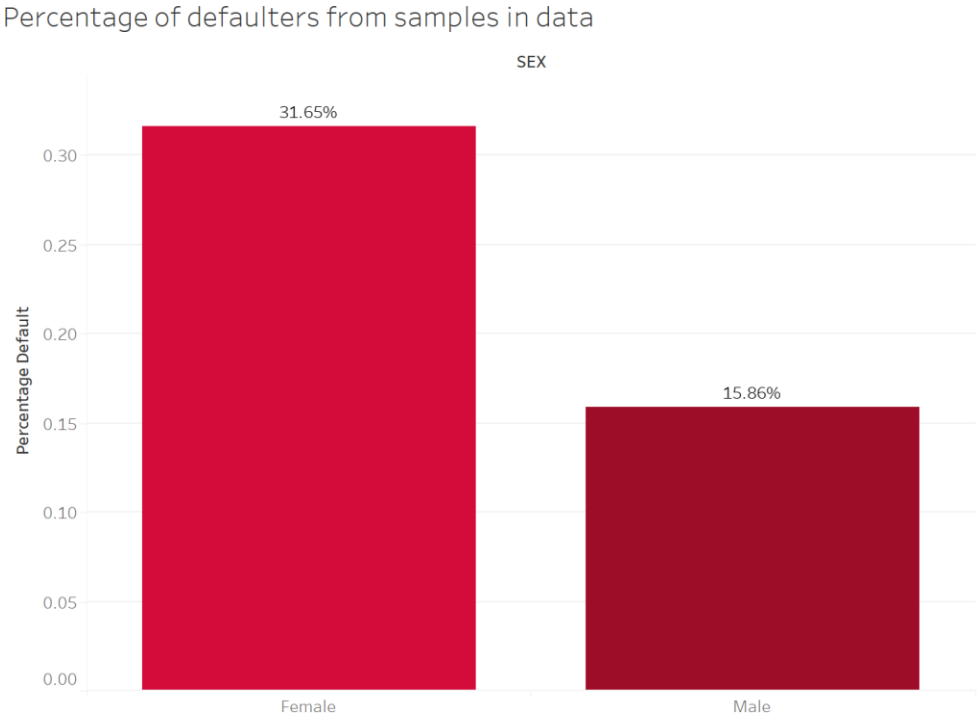
Replaced invalid values in MARRIAGE with 'others'

Replaced invalid values in EDUCATION with 'others'

Detected negative values in Bill Statement and Payment amounts but chose to ignore due to large number of such values in all columns

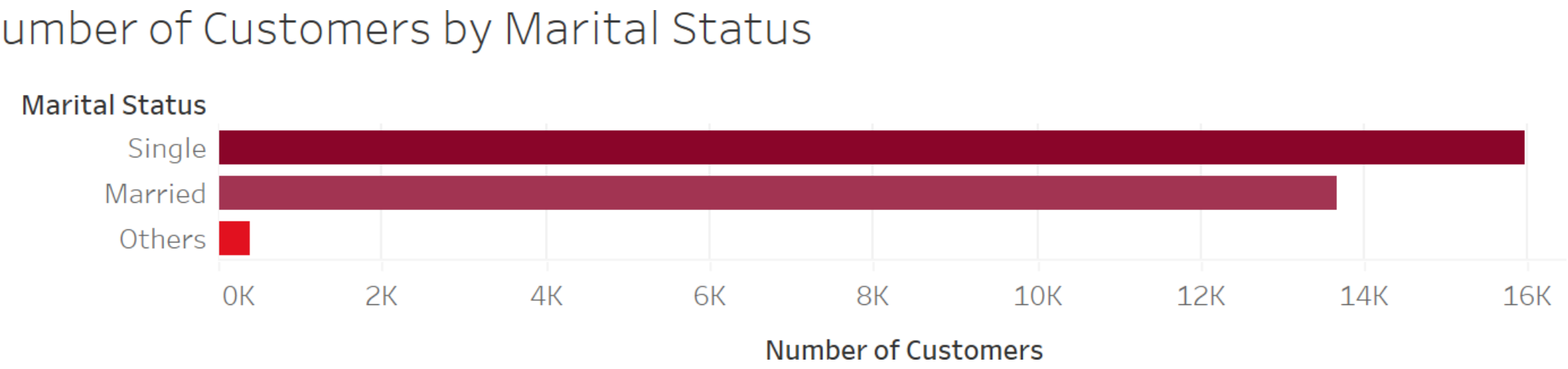
DATA CLEANING: INVALID AND NEGATIVE VALUES





MAPPING 'SEX' WITH DEFAULTS

GREATER NUMBER OF FEMALE DEFAULTERS

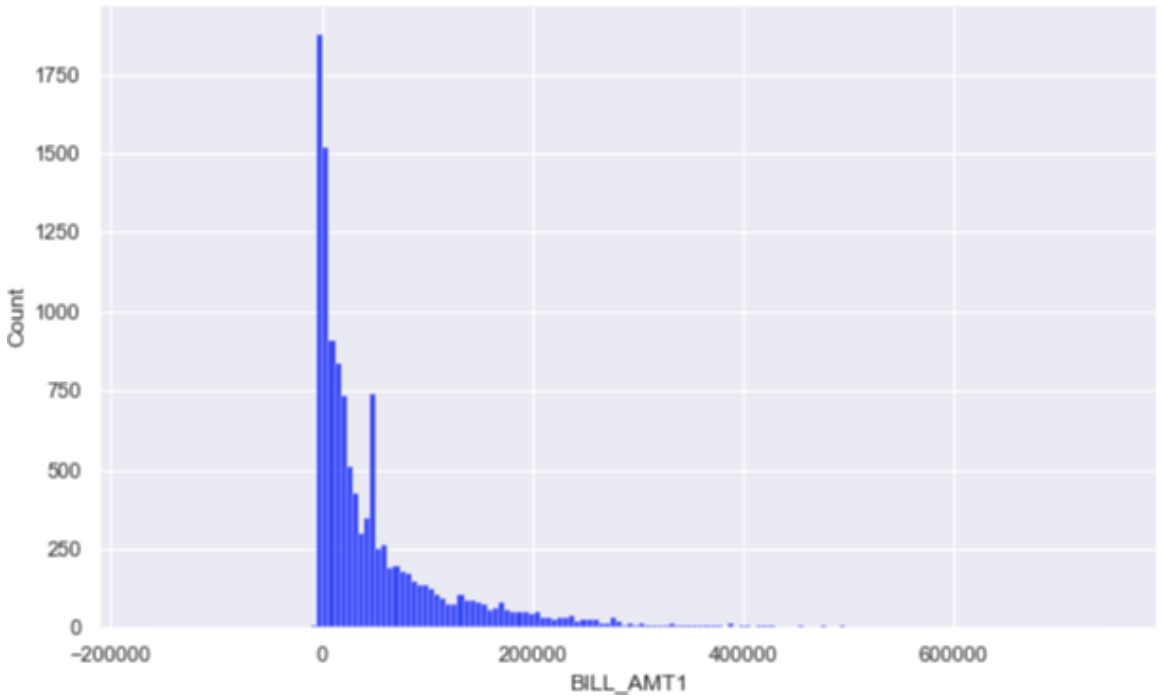


MARITAL STATUS OF CUSTOMERS

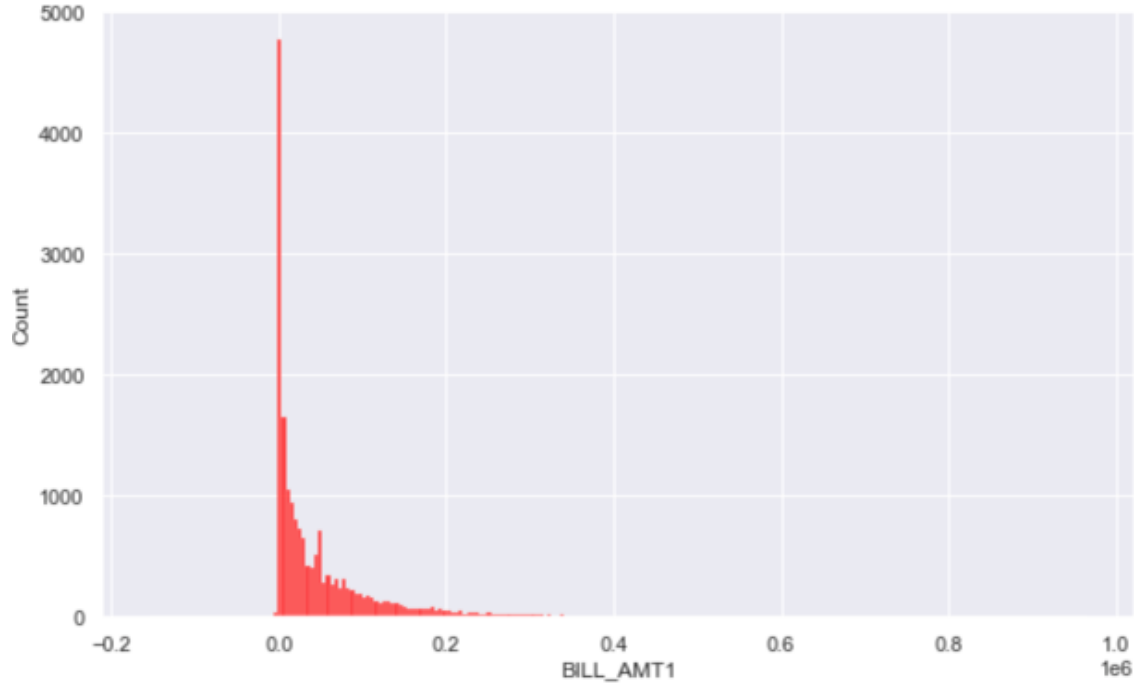
GREATEST NUMBER OF SINGLE CUSTOMERS

MAPPING 'SEX' WITH BILL AMOUNT 1

DISTRIBUTION OF BILL AMOUNT 1 IS SIMILAR ACROSS BOTH MALE & FEMALE



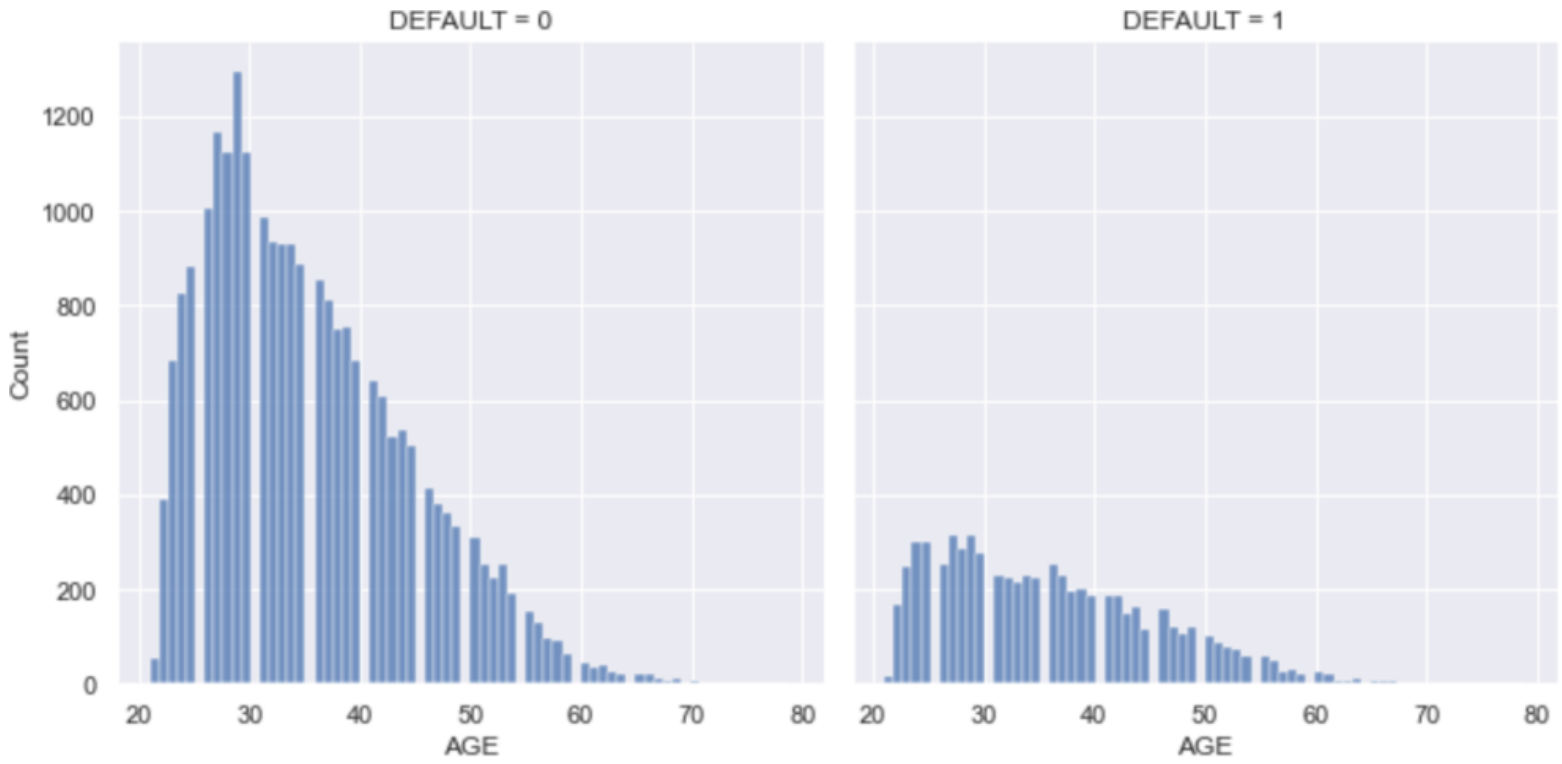
Females

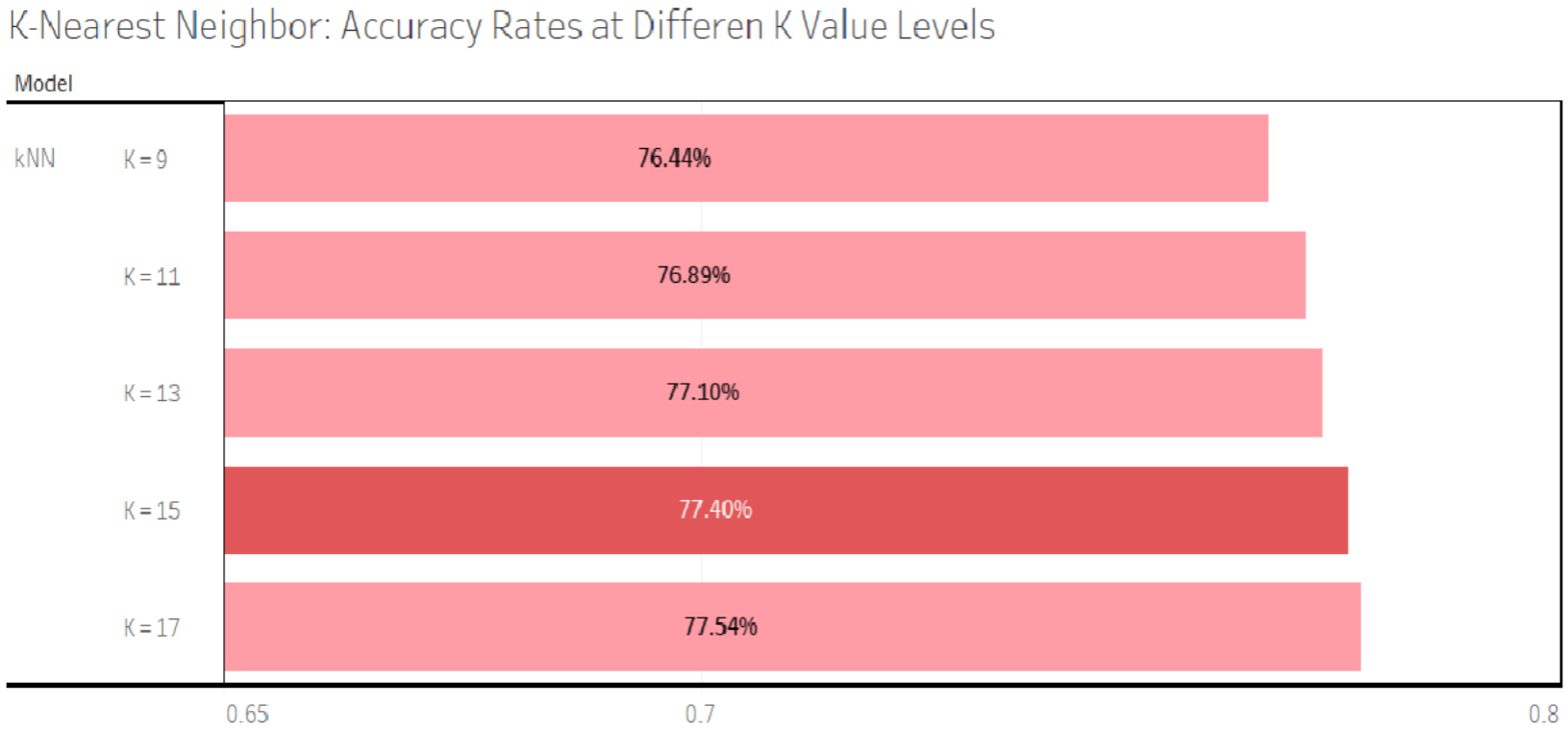


Males

A RELATIONSHIP BETWEEN AGE AND DEFAULTS

- No linear or direct relationship between age and Defaults
- Max defaulters between ages 25 and 30
- After 30, defaults exhibit a decreasing pattern

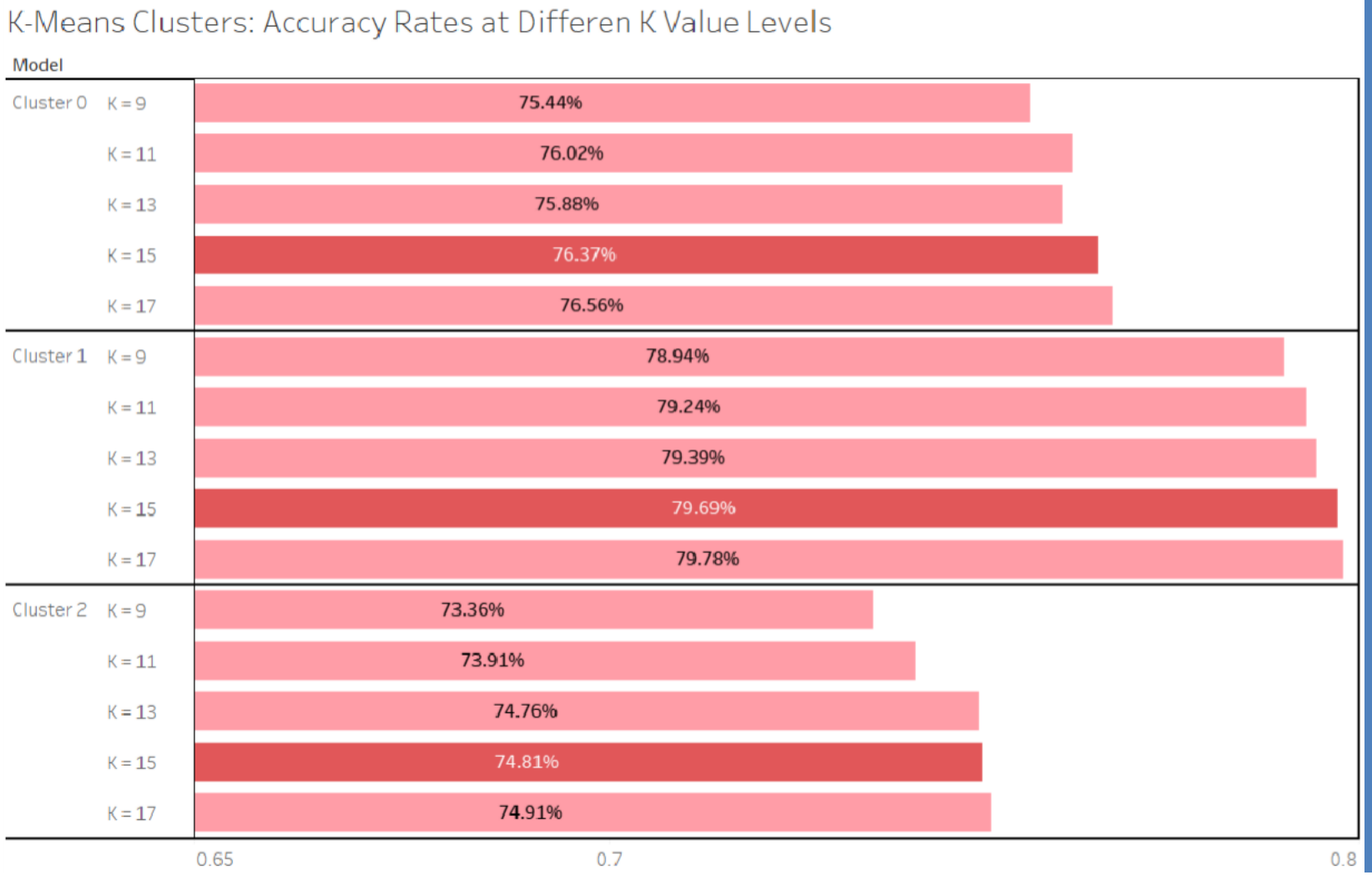




MAPPING K-VALUES FOR K-NEAREST NEIGHBOR MODEL

We chose a K-Value of 15 because:

- Accuracy Rate beyond K=15 increases insignificantly



MAPPING K-VALUES FOR K-MEANS CLUSTERS MODELS

We chose a K-Value of 15 because:

- Accuracy Rate beyond K=15 increases insignificantly

Neural Network Comparisons

Model1	Accuracy	True Positive Rate	True Positive (Count)
NN (1,1)	81.98%	36.23%	721
NN (1,2)	81.68%	31.86%	634
NN (1,3)	82.17%	38.29%	762
NN (1,4)	82.10%	38.14%	759
NN (1,5)	81.84%	34.52%	687
NN (2,1)	81.90%	36.71%	740
NN (2,2)	82.07%	36.56%	737
NN (2,3)	82.07%	33.98%	685
NN (2,4)	81.34%	31.66%	630
NN (2,5)	81.50%	32.56%	648

CHOOSING OPTIMAL NEURAL NETWORK MODEL

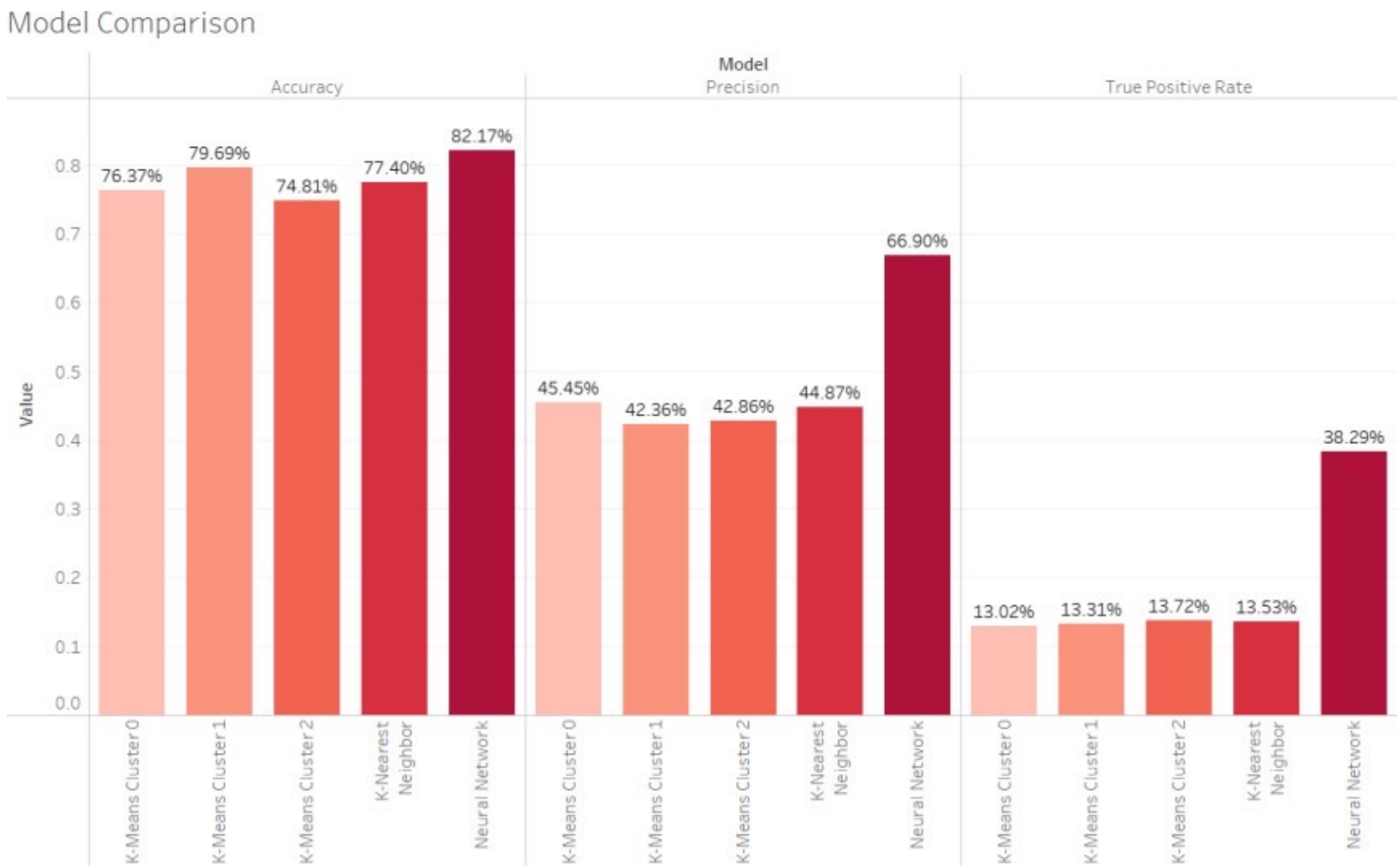
We chose 1 hidden layer with 3 hidden neurons per layer because:

- Highest Accuracy Rate (82.10%)
- Highest True Positive Rate (38.29%)
- Highest True Positive Count (762)

MODEL BASED ON **NEURAL NETWORK** ALGORITHM SIGNIFICANTLY OUTPERFORMS OTHERS

Stronger in 3 Quantifiable Metrics:

- Accuracy Rate
- Precision
- True Positive Rate



CONCLUSION

Recommendations:

- Use a Neural Network model to get the best accuracy
  - Outperforms other models in 3 metrics
  - Utilizes a full, non-segmented training set
- Reevaluate the model after 6 months for further understanding of predictive value of the Neural Networking algorithm to the company

APPENDIX
NUMBER OF CUSTOMERS IN DATASET (Q 1.1)

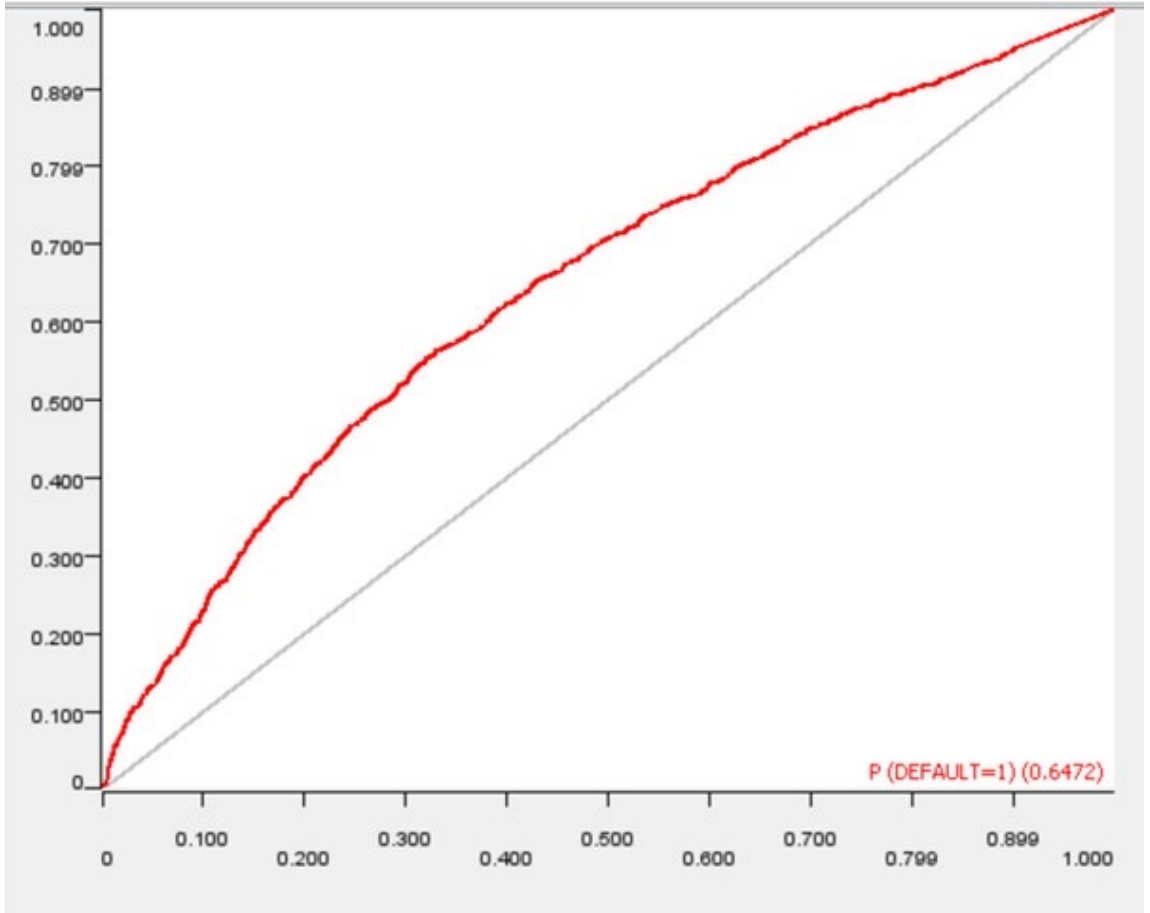
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  int64
 2   SEX                         30000 non-null  int64
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  int64
 5   AGE                         30000 non-null  int64
 6   PAY_0                       30000 non-null  int64
 7   PAY_2                       30000 non-null  int64
 8   PAY_3                       30000 non-null  int64
 9   PAY_4                       30000 non-null  int64
 10  PAY_5                       30000 non-null  int64
 11  PAY_6                       30000 non-null  int64
 12  BILL_AMT1                   30000 non-null  int64
 13  BILL_AMT2                   30000 non-null  int64
 14  BILL_AMT3                   30000 non-null  int64
 15  BILL_AMT4                   30000 non-null  int64
 16  BILL_AMT5                   30000 non-null  int64
 17  BILL_AMT6                   30000 non-null  int64
 18  PAY_AMT1                    30000 non-null  int64
 19  PAY_AMT2                    30000 non-null  int64
 20  PAY_AMT3                    30000 non-null  int64
 21  PAY_AMT4                    30000 non-null  int64
 22  PAY_AMT5                    30000 non-null  int64
 23  PAY_AMT6                    30000 non-null  int64
 24  default payment next month  30000 non-null  int64
dtypes: int64(25)
memory usage: 5.7 MB
```

There are 30,000 customers in the sample

APPENDIX (CONT.)
KNN MODEL CONFUSION MATRIX, ROC CURVE AND AUC (Q 3.2)

DEFAULT \...	1	0
1	236	1770
0	292	6702

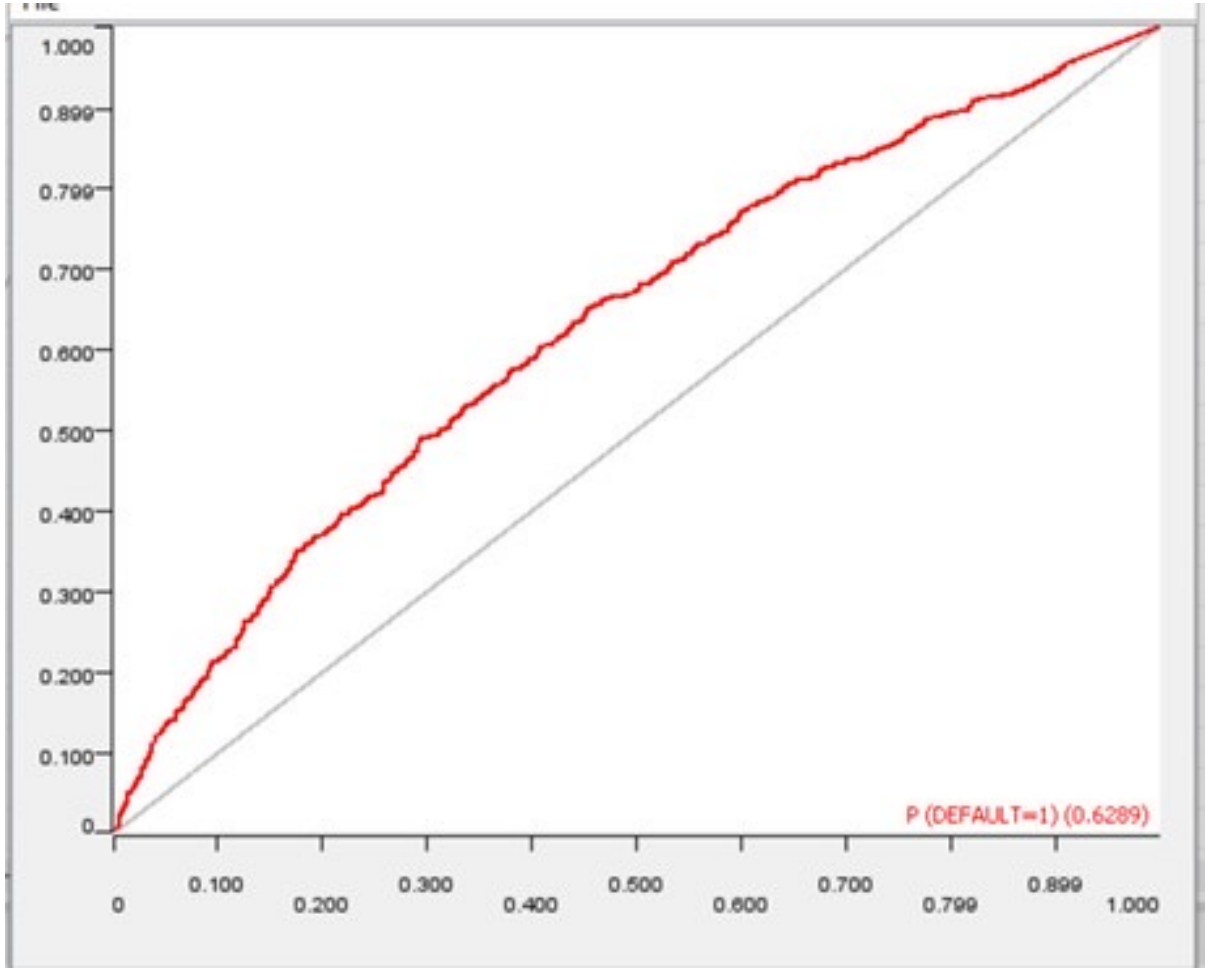
Area Under Curve : 0.6472 sq. units



APPENDIX (CONT.)
K-MEANS CLUSTER 0 CONFUSION MATRIX, ROC CURVE AND AUC (Q'S 3.6 & 3.8)

DEFAULT \...	1	0
1	104	676
0	145	2744

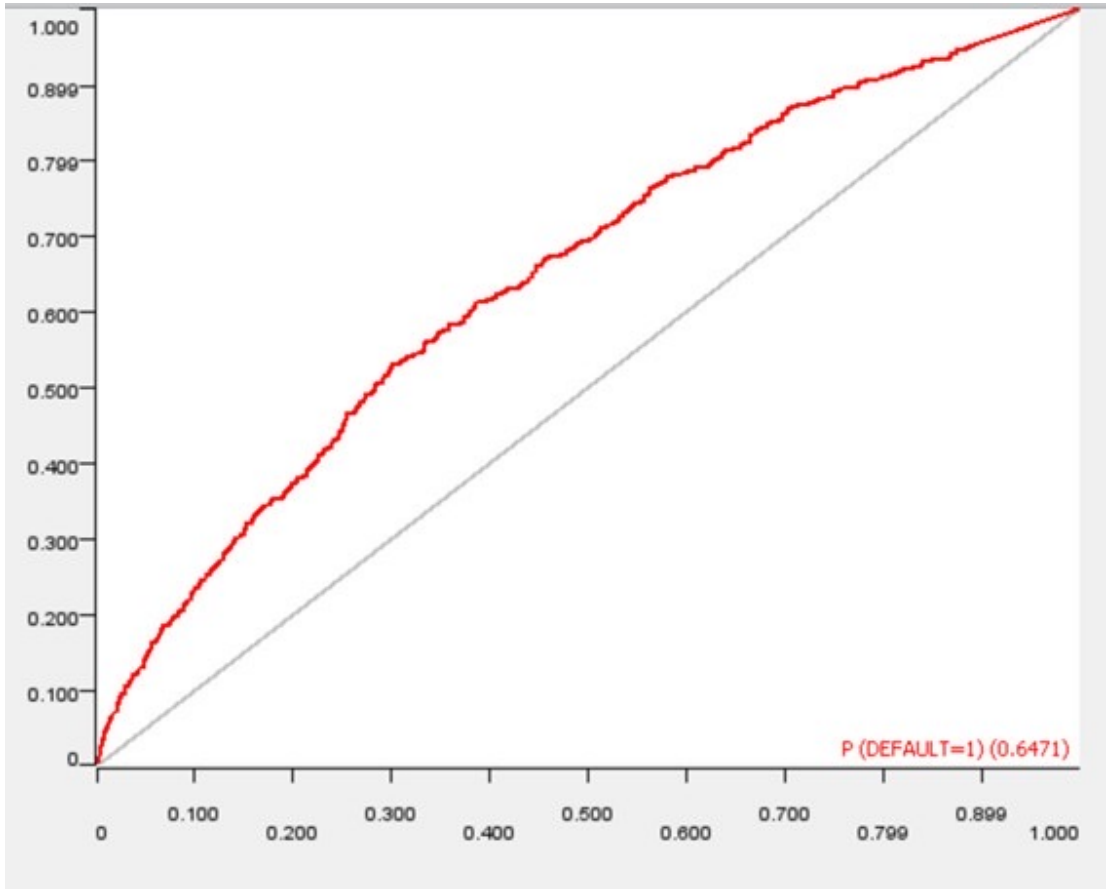
Area Under Curve : 0.6289 sq. units



APENDIX (CONT.)
K-MEANS CLUSTER 1 CONFUSION MATRIX, ROC CURVE AND AUC (Q'S 3.6 & 3.8)

DEFAULT \...	1	0
1	87	613
0	111	2523

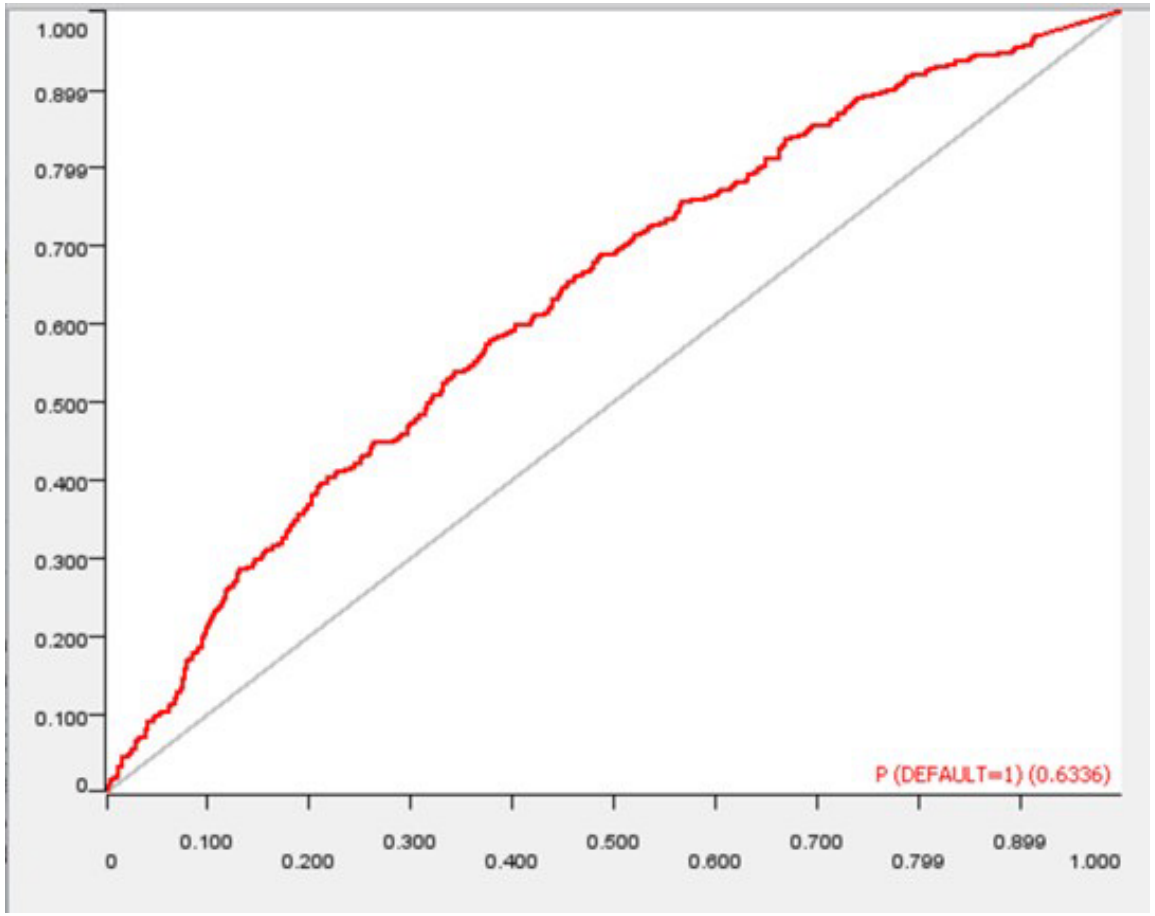
Area Under Curve : 0.6471 sq. units



APENDIX (CONT.)
K-MEANS CLUSTER 2 CONFUSION MATRIX, ROC CURVE AND AUC (Q'S 3.6 & 3.8)

DEFAULT \...	1	0
1	57	430
0	100	1410

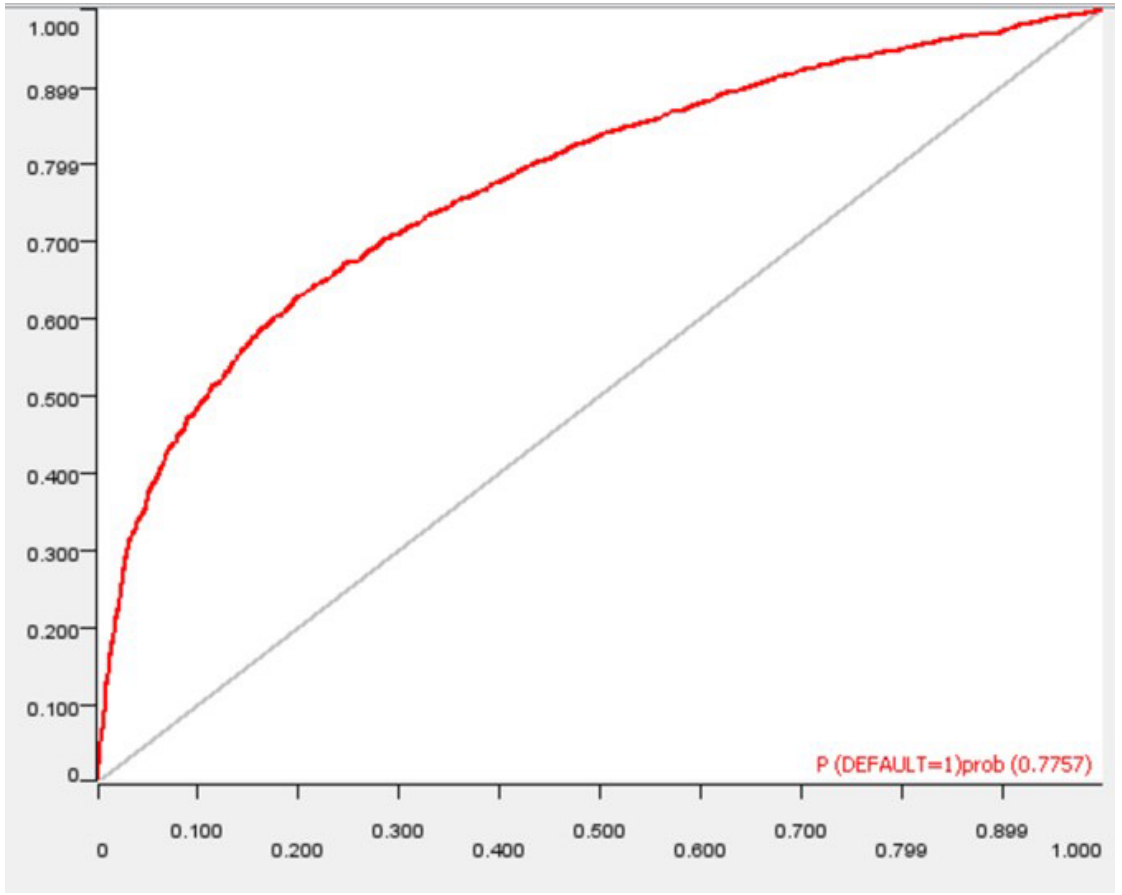
Area Under Curve : 0.6336 sq. units



APENDIX (CONT.)
NEURAL NETWORK CONFUSION MATRIX, ROC CURVE AND AUC (Q 4.2)

Row ID	1	0
1	704	1312
0	321	6663

Area Under Curve : 0.7758 sq. units



APENDIX (CONT.)
K-MEANS VS. K-NEAREST NEIGHBOR CLASSIFICATION COMPARISON (Q 3.9)

K-Means / K-Nearest Neighbor Classification Comparison

Model	Accuracy (%)	Misclassification (%)	True Positve	False Positve	True Negative	False Negative	Total Sample Size
K-Means Cluster 0	76.37%	23.63%	110	132	2,692	735	3,669
K-Means Cluster 1	79.69%	20.31%	86	117	2,571	560	3,334
K-Means Cluster 2	74.81%	25.19%	66	88	1,428	415	1,997
K-Nearest Neighbor	77.40%	22.60%	267	328	6,699	1,706	9,000

K-Means Cluster 2 outperforms the non-segmented K-Nearest Neighbor:

- Higher Accuracy Rate
- Low Misclassification Rate

APENDIX (CONT.)
COMPARISON OF METRICS FROM ALL MODELS (Q'S 3.3, 3.7, 4.3, & 5.1)

Model	Accuracy	Misclassification	False Positve Rate	True Positive Rate	Specificity	Precision	Prevalence
K-Means Cluster 0	76.37%	23.63%	4.67%	13.02%	95.33%	45.45%	23.03%
K-Means Cluster 1	79.69%	20.31%	4.35%	13.31%	95.65%	42.36%	19.38%
K-Means Cluster 2	74.81%	25.19%	5.80%	13.72%	94.20%	42.86%	24.09%
K-Nearest Neighbor	77.40%	22.60%	4.67%	13.53%	95.33%	44.87%	21.92%
Neural Network	82.17%	17.83%	5.38%	38.29%	94.62%	66.90%	22.11%

Why we chose Neural Network :

- Highest accuracy
- Highest True Positive Rate
- Least Misclassification Rate (100 – Accuracy)
- Highest Precision
- 2nd Highest Specificity