Group 23

Loan Default Prediction

Based on historical data, we built a set of predictive models to forecast loan default for customers based on their profile. We also investigated how different factors (age, marital status, etc.) could affect the possibility of loan default

Chi
David
Yin

Executive Summary

1

ANN model is the best performing predictive model

Through simulation, ANN model has the strongest ability to predict loan default with **AUC 76.31%** and **81.78% accuracy**

2

Males and the married are more likely to default loans

Female have more default cases than males, but **males have higher loan default probability of 31.87%** (female 26.2%). The **married default rate is 30.67%**, while for 26.47% singles.

3

Recommendation

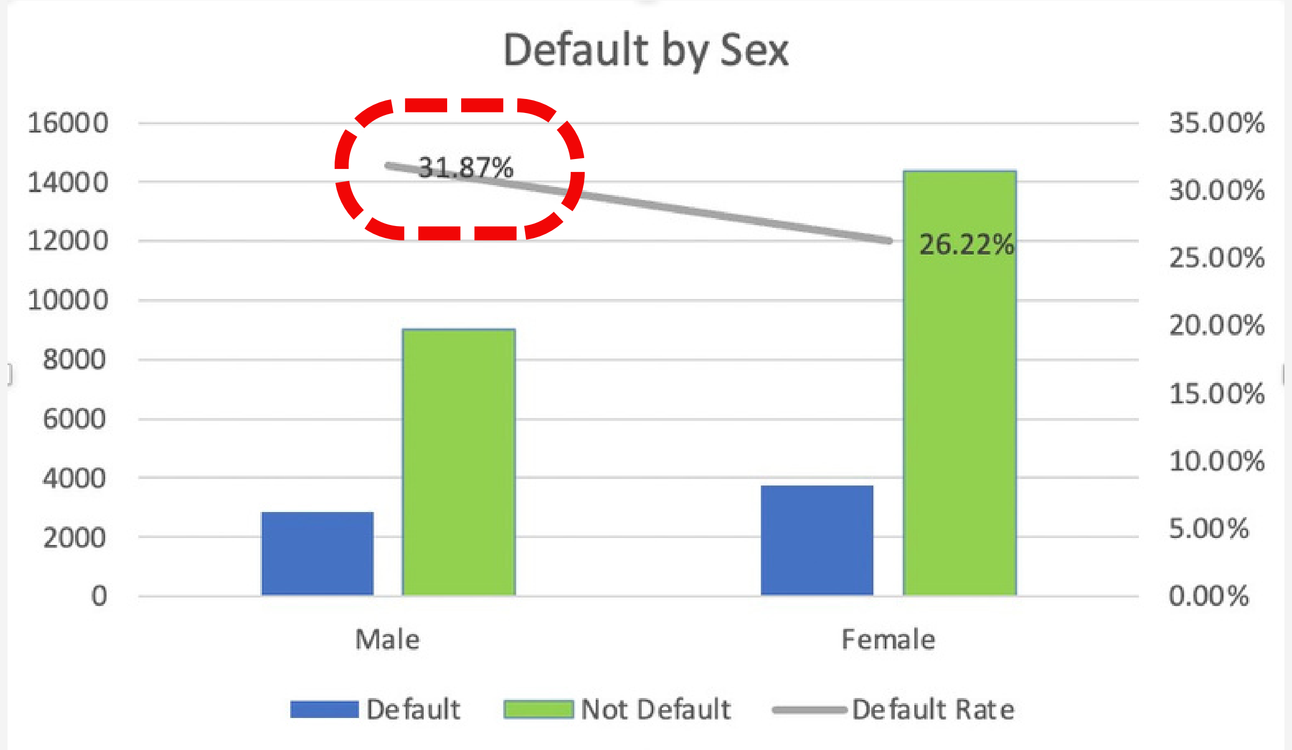
- Adopt best performing **ANN model**
- Adjust **higher rates** for customers more likely to default
- **Grant small amount loans** rather than big amount
- Establish **creditscore** for credit card clients

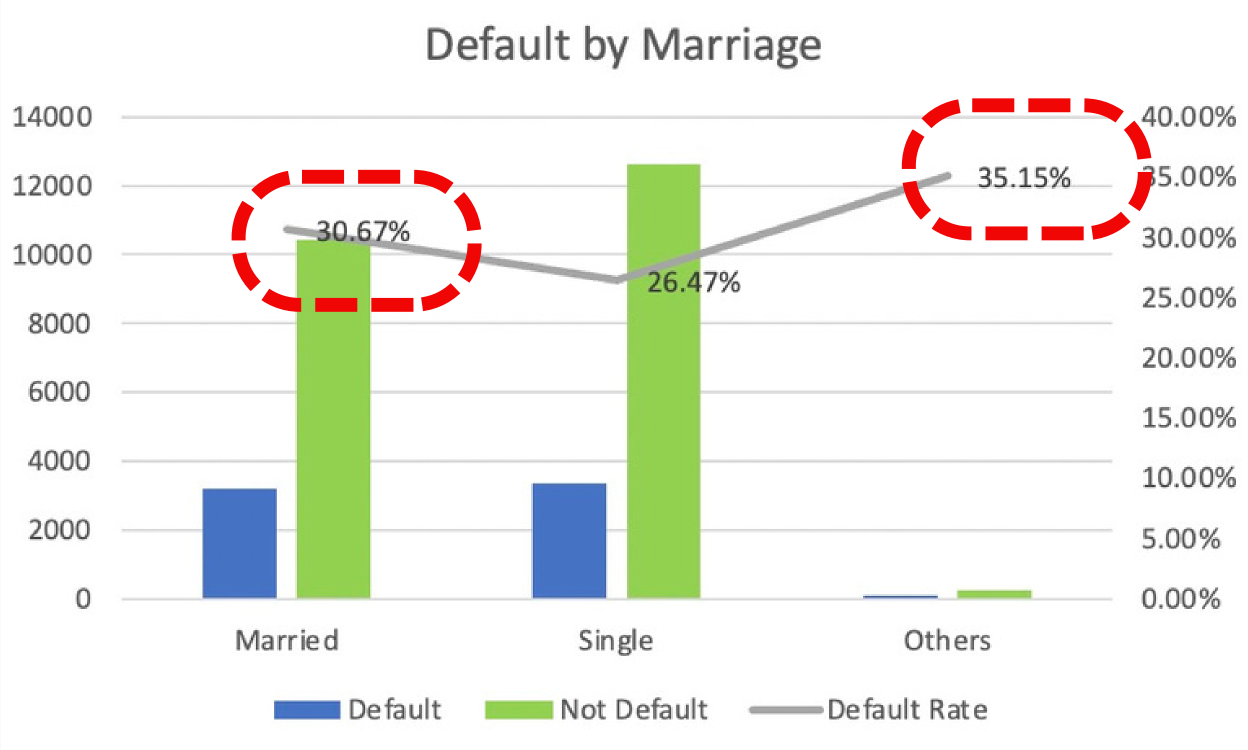
Agenda

- **Understanding data -** Variables, EAD
- **Preprocessing** - Encode variables
- **Model Exploration -** Build and evaluate models
- **Model Comparison -** Find the best predicting model
- **Recommendation -** More profit, lower default risk

Understanding Data

Customers of different **sex** and **marital status** have varying default rates

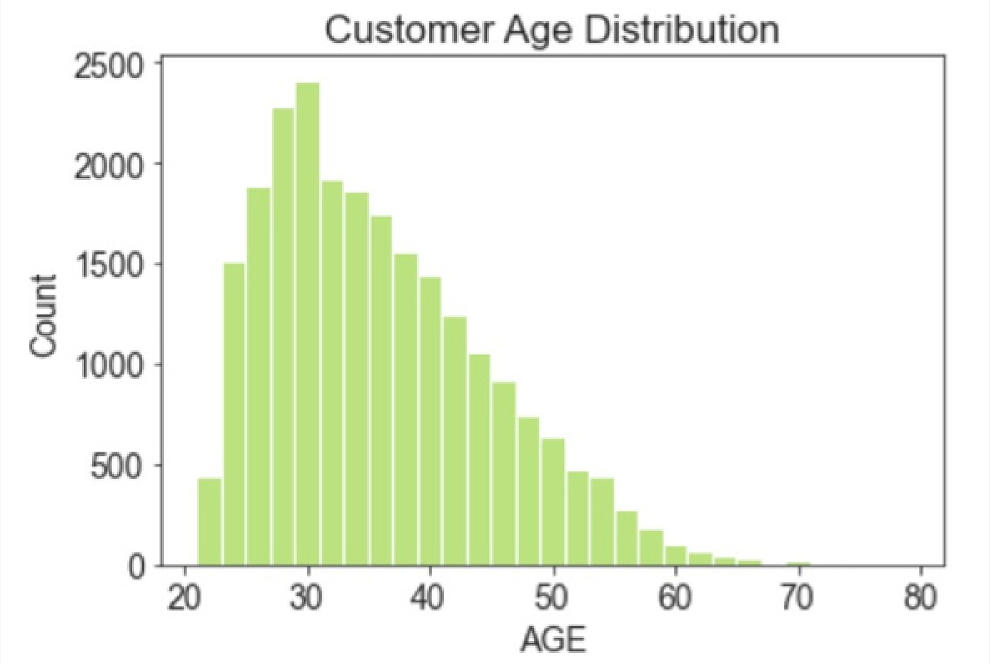


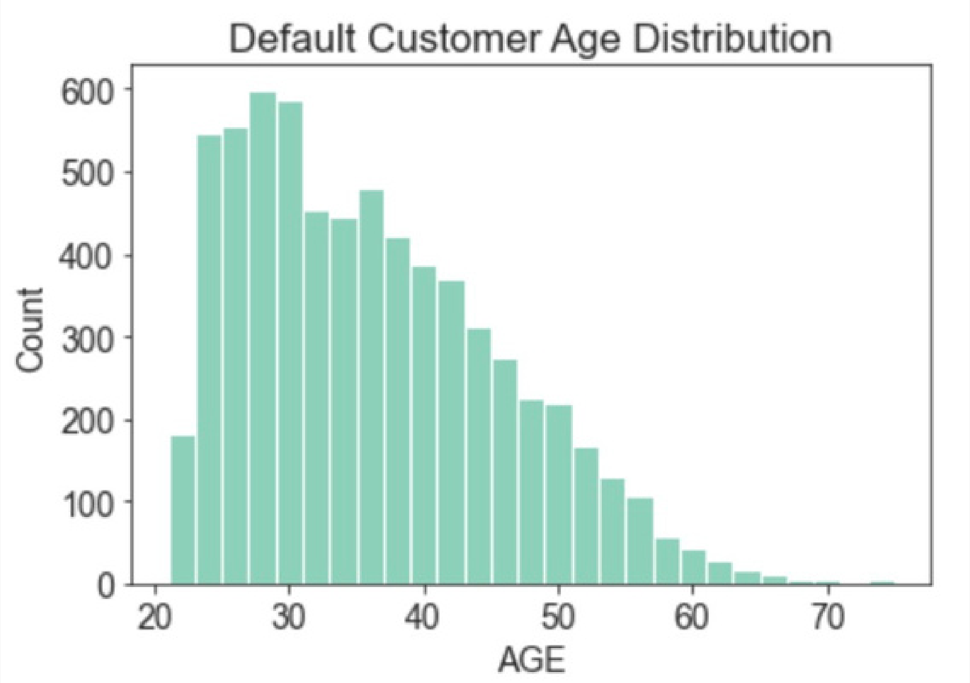


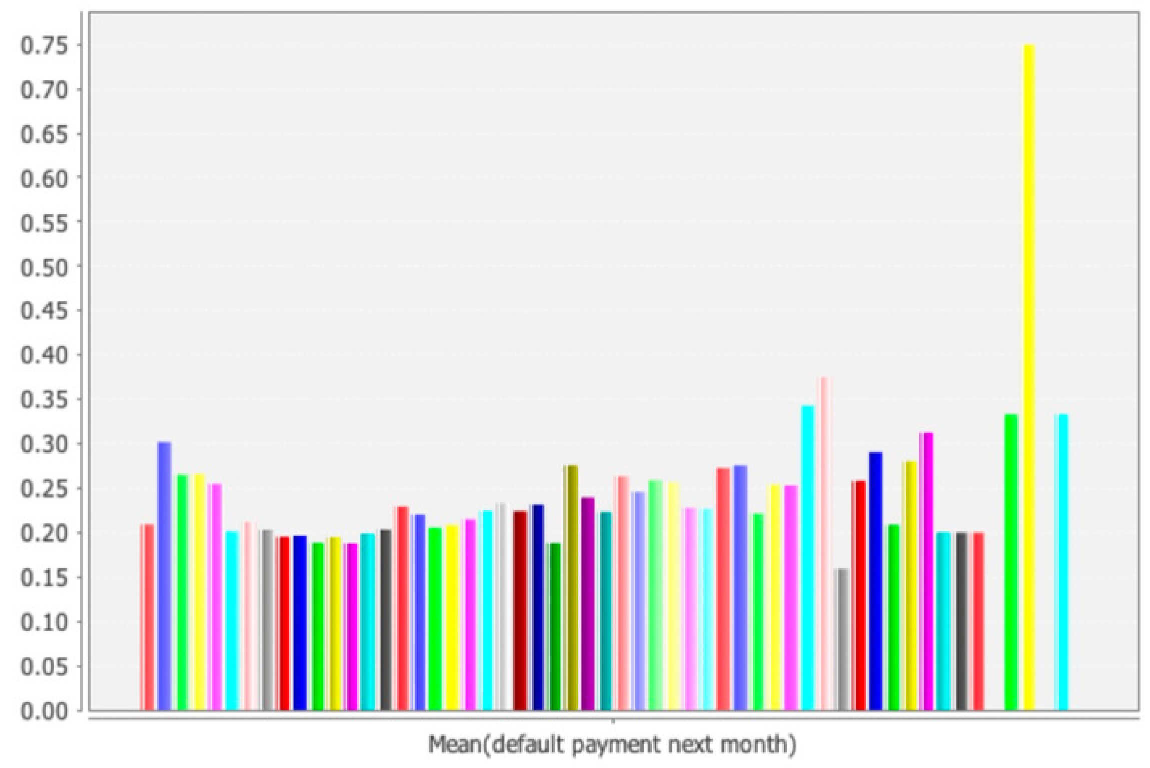
- There are **30,000 observations** and **25 variables** in the raw data
- **Female** has the most defaults
- **Male** has higher possibility to default
- **Married** and **other marrital status** more likely to default

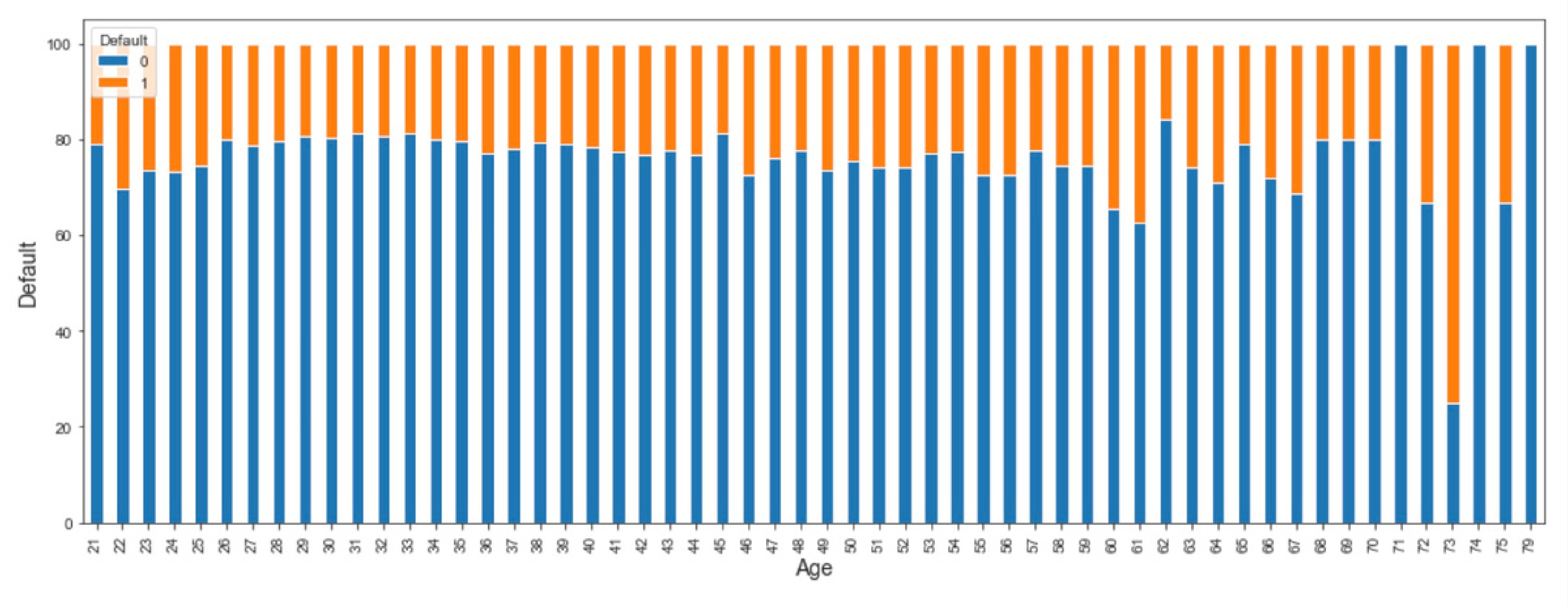
Understanding Data

Age has **no correlation** with loan defaults





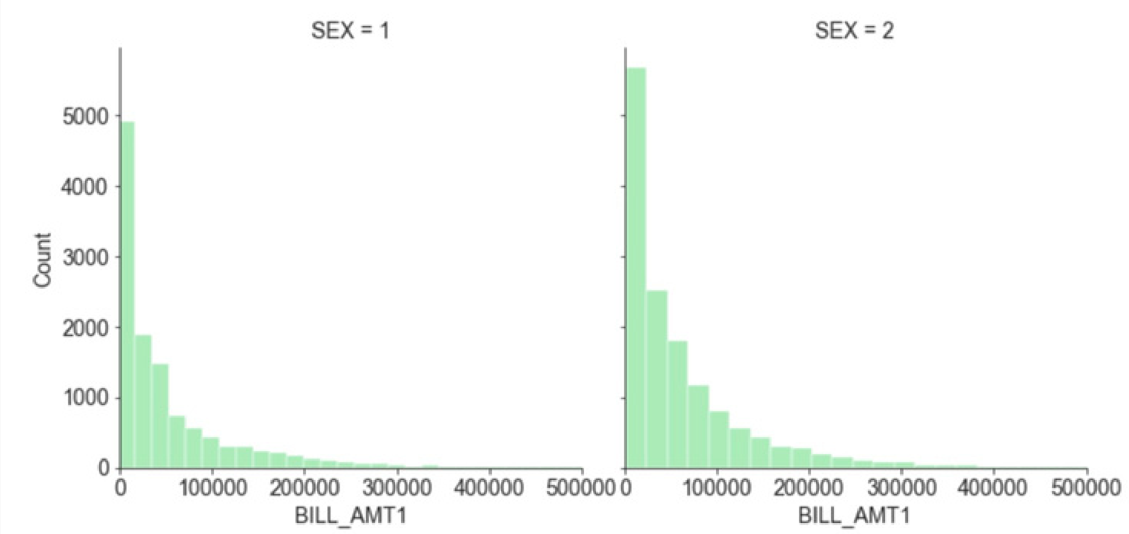


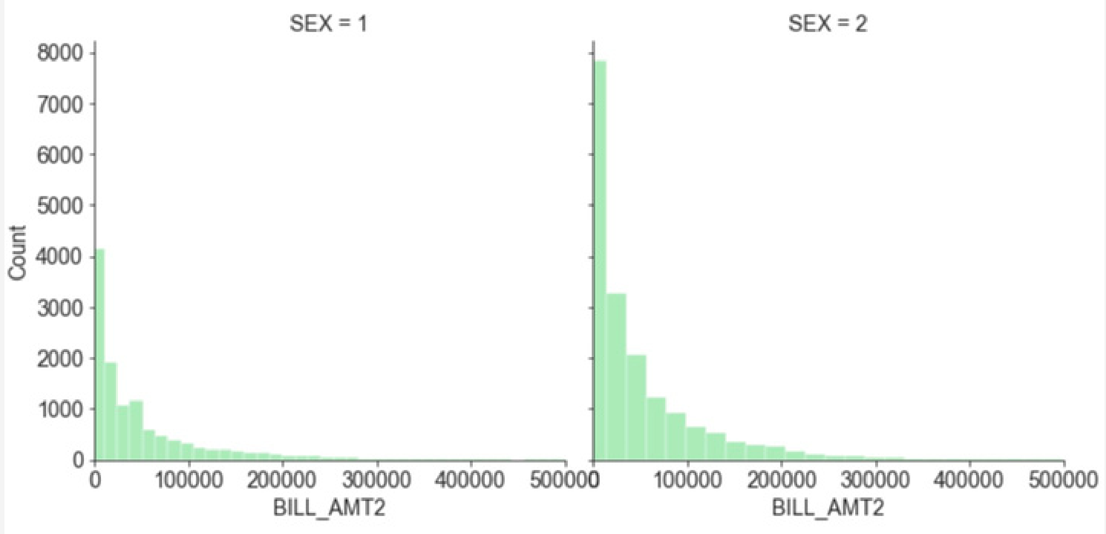


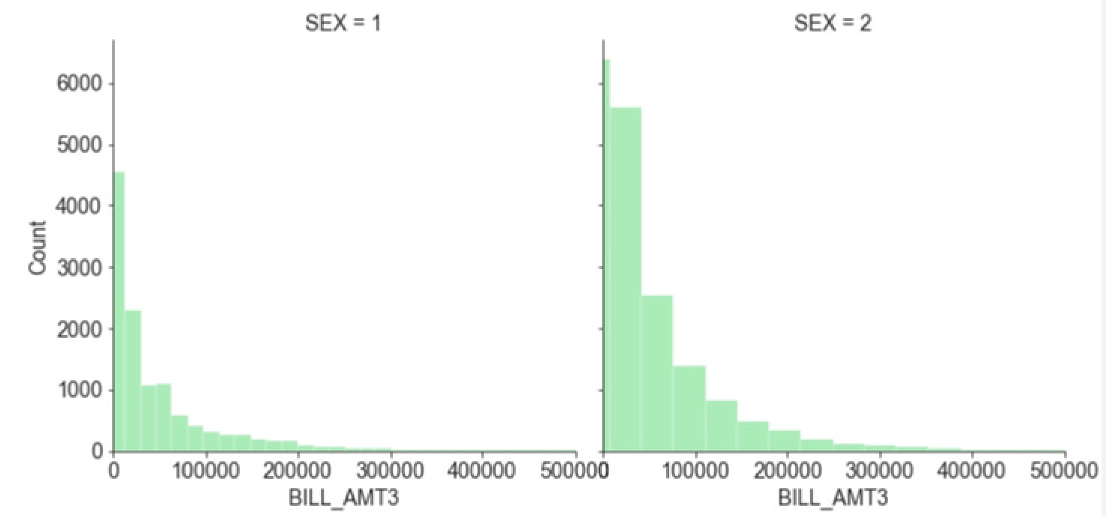
- Statistically, Pearsons correlation is **low** at 0.014
- Graphically, default probability is not related to age

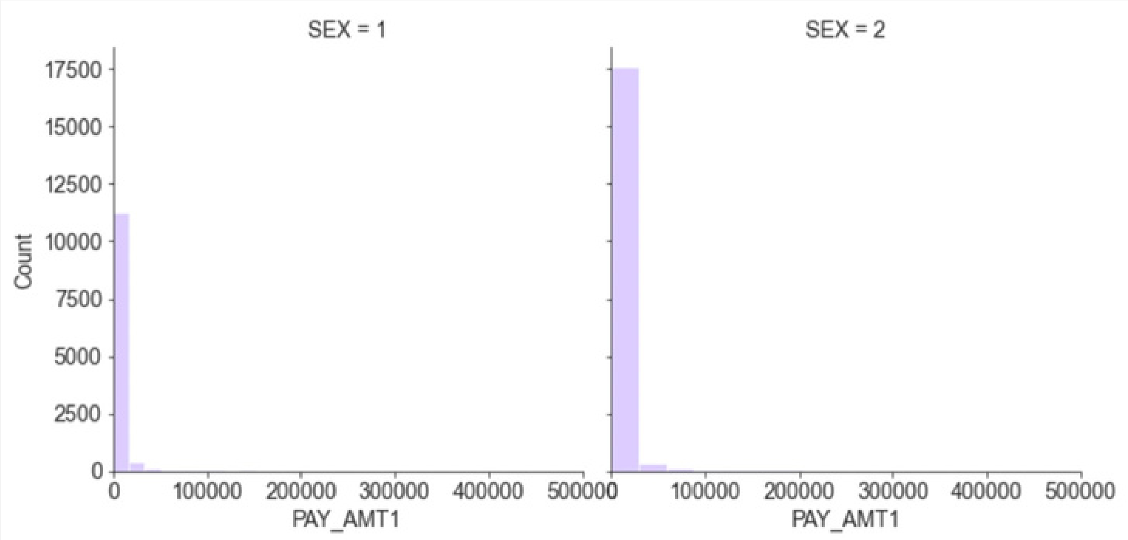
Understanding Data

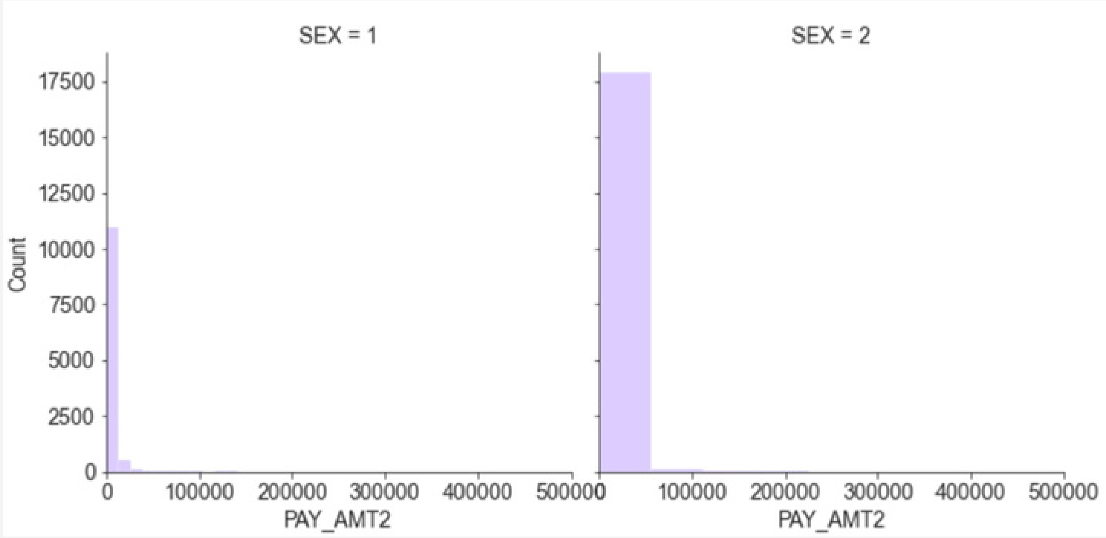
Customers with large amount bills, no matter males or females, they are **paying less** than they are supposed to.

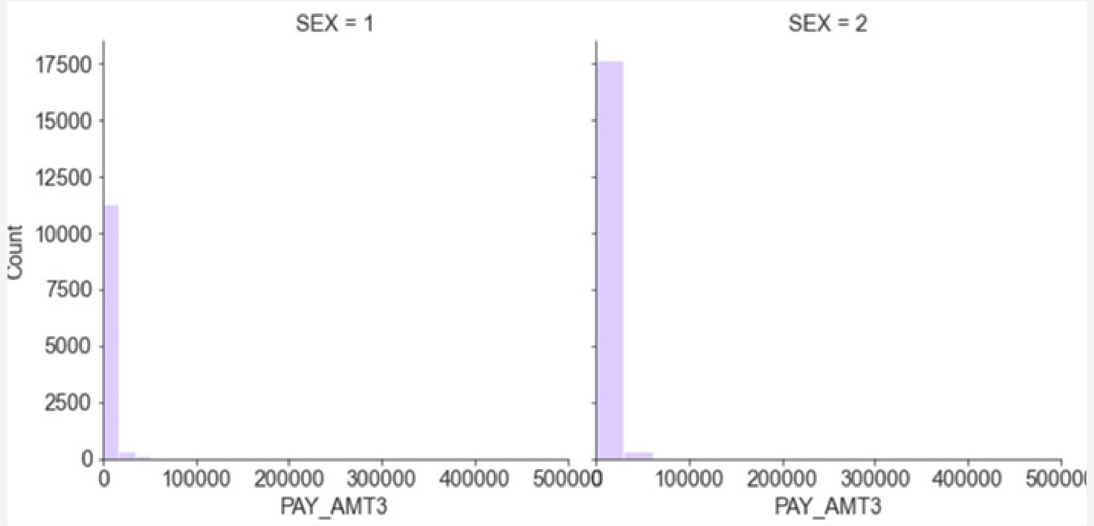












Preprocessing: Improve Model Efficiency

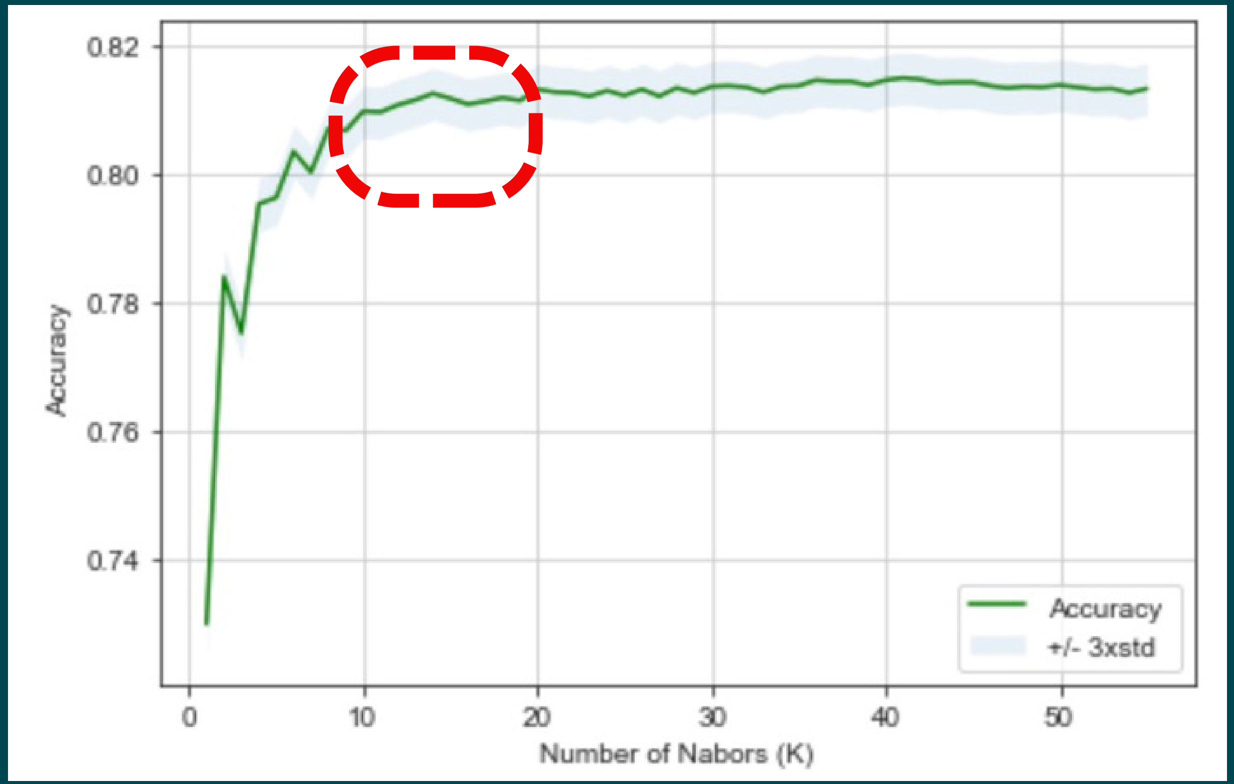
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 24 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   LIMIT_BAL  30000 non-null  int64
 1   SEX        30000 non-null  category
 2   EDUCATION  30000 non-null  category
 3   MARRIAGE   30000 non-null  category
 4   AGE        30000 non-null  int64
 5   PAY_1      30000 non-null  category
 6   PAY_2      30000 non-null  category
 7   PAY_3      30000 non-null  category
 8   PAY_4      30000 non-null  category
 9   PAY_5      30000 non-null  category
 10  PAY_6      30000 non-null  category
 11  BILL_AMT1  30000 non-null  int64
 12  BILL_AMT2  30000 non-null  int64
 13  BILL_AMT3  30000 non-null  int64
 14  BILL_AMT4  30000 non-null  int64
 15  BILL_AMT5  30000 non-null  int64
 16  BILL_AMT6  30000 non-null  int64
 17  PAY_AMT1   30000 non-null  int64
 18  PAY_AMT2   30000 non-null  int64
 19  PAY_AMT3   30000 non-null  int64
 20  PAY_AMT4   30000 non-null  int64
 21  PAY_AMT5   30000 non-null  int64
 22  PAY_AMT6   30000 non-null  int64
 23  Default    30000 non-null  category
dtypes: category(10), int64(14)
```

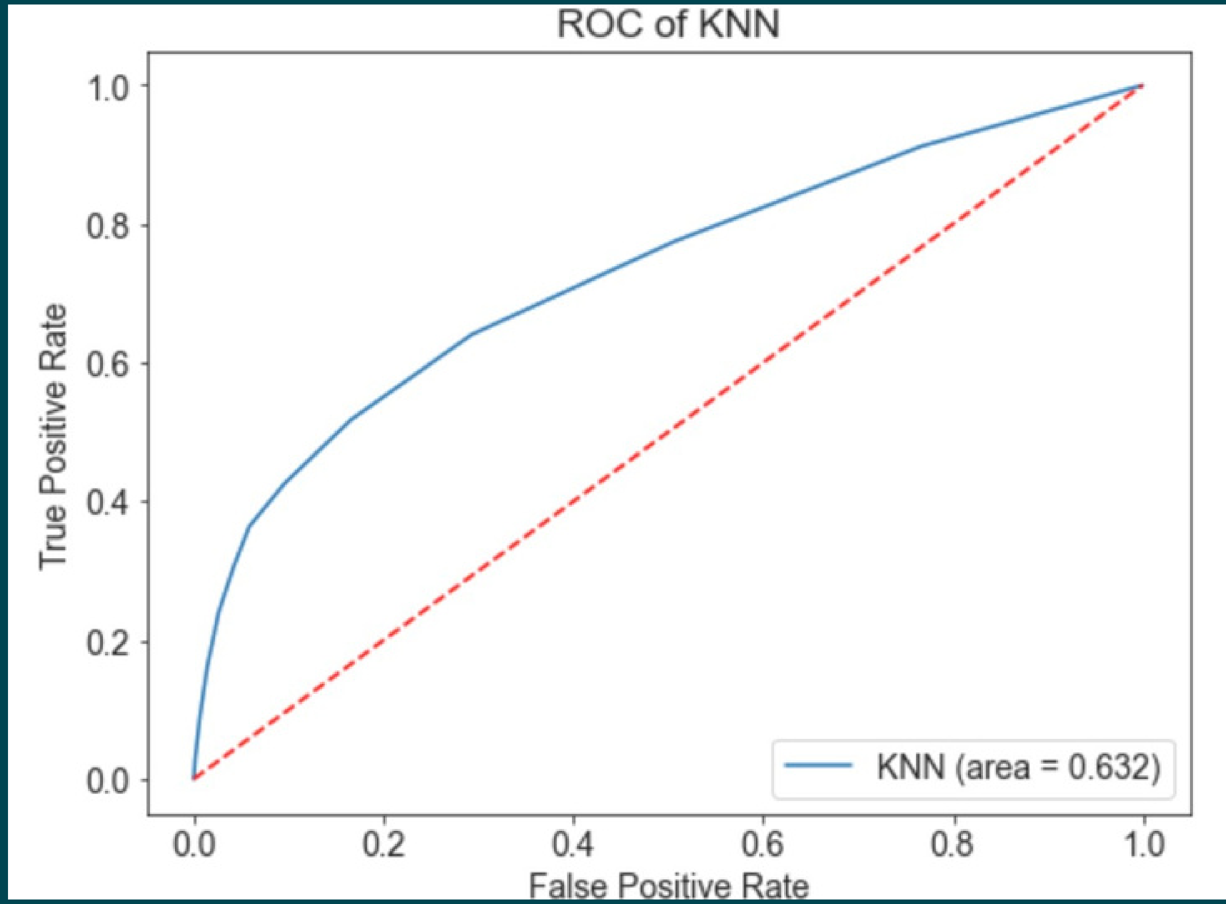
- Keep negative bill amounts
- No Missing Values
- leave the outliers

- **Categorical Variables:** Data is in numeric form, we use OneHot encoding to create binary dummy variables eliminating the affect that number order bring to the models
- **Numeric Variables:** Normalization the numbers (scale the variables to similar sizes) makes the graphing of the values more efficient for the ML methods

KNN Model

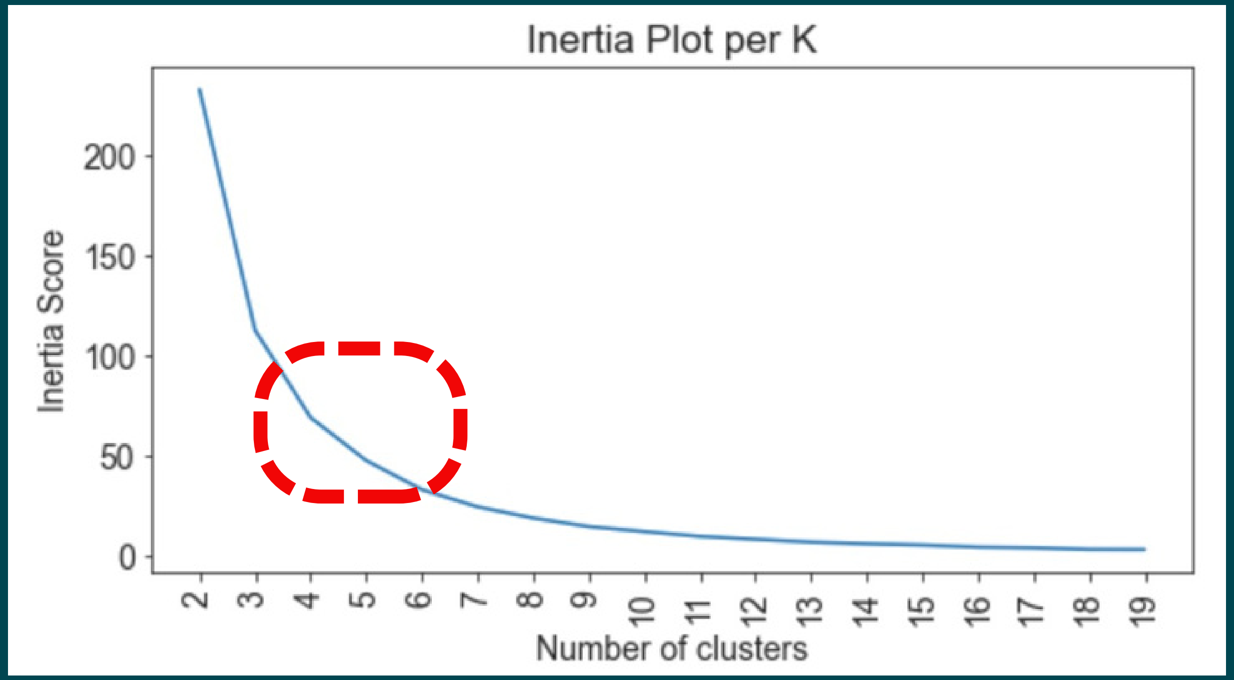
- Best result with 12 class
- Ability to predict that a client defaults next month: 63.22%
- Balance tradeoff of variability and bias
- Performed with 81.08% accuracy

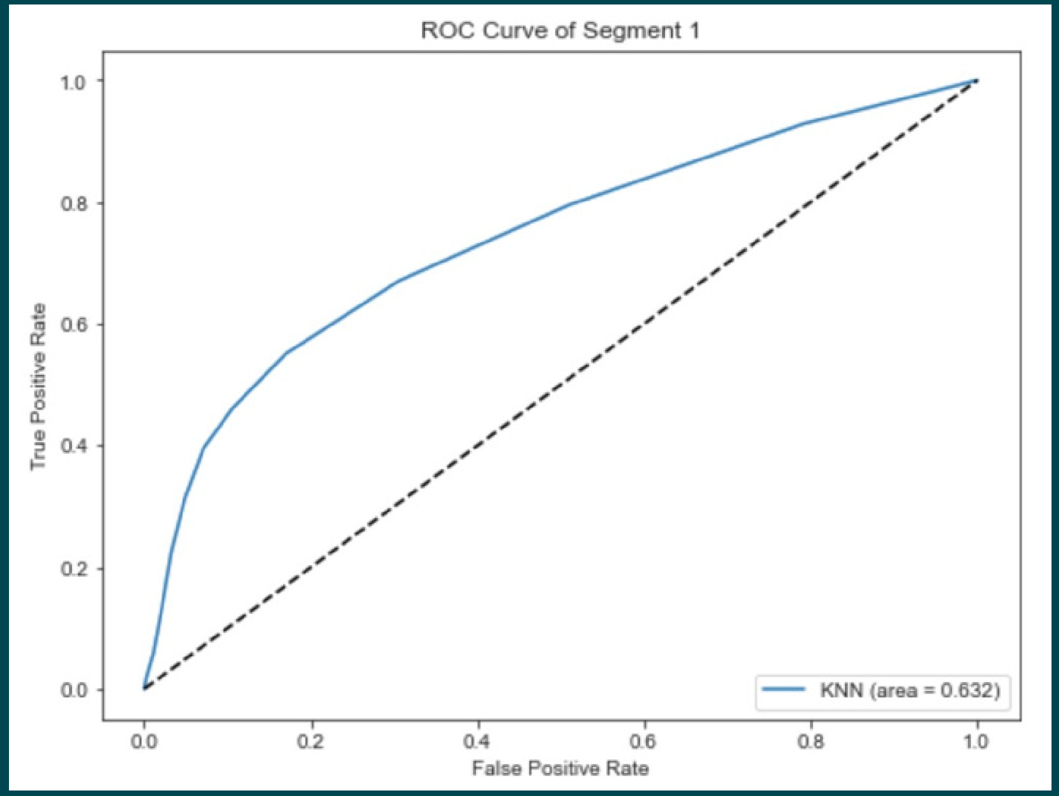




K-means Cluster Model

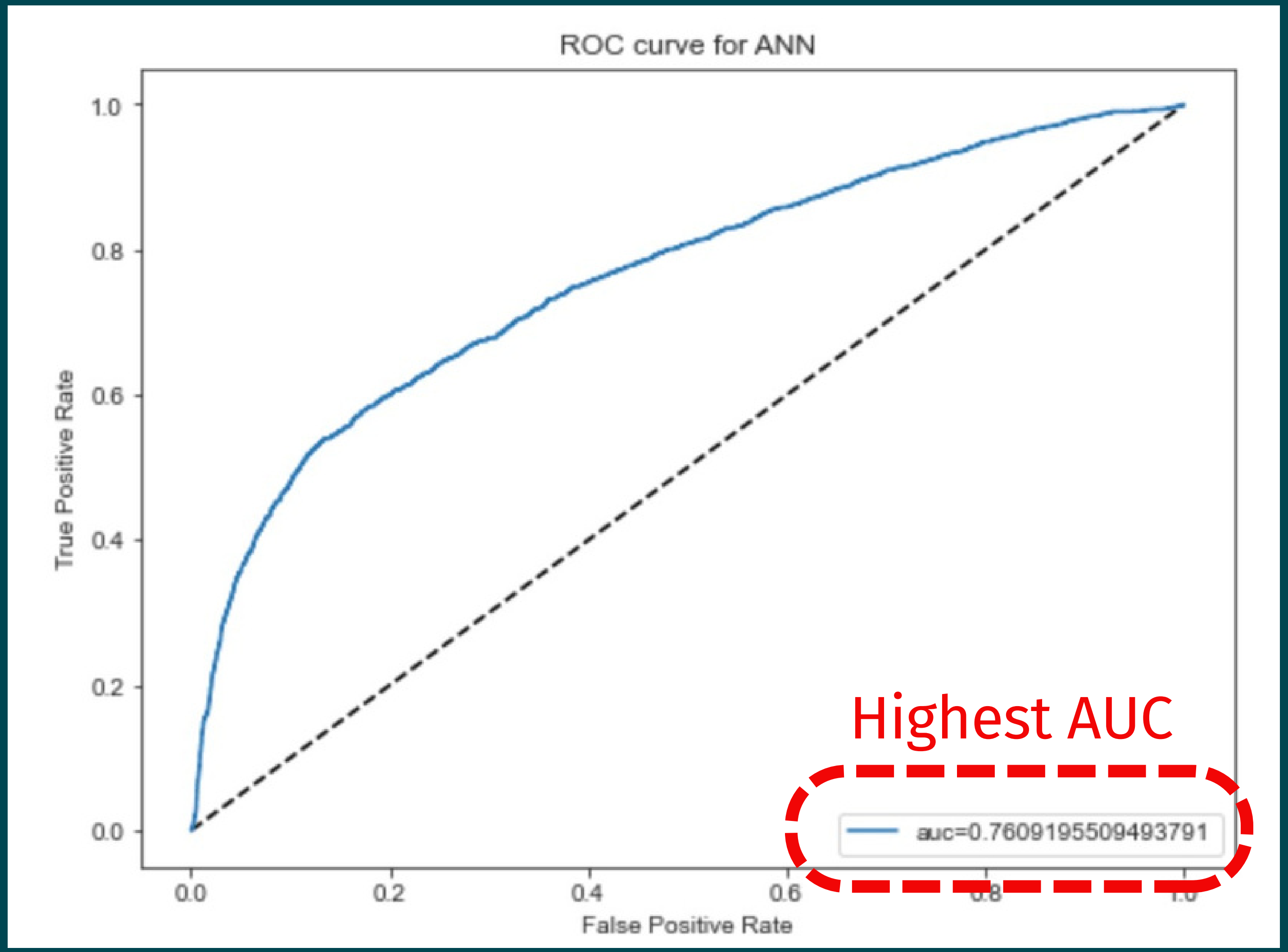
- Segmented age into 4 clusters
- Best group's ability to predict a client defaults next month is 63.17%
- Best group performed with 80.81% accuracy
- Other group performed with accuracy 77.04%, 82.60%,80.72%
- No group has better classification performance than the non-segmented KNN model (comparing AUC)





ANN Model

- The best performing model
- Ability to predict that a client defaults next month: 76.31%
- Performed with 81.78% accuracy



Recommendation

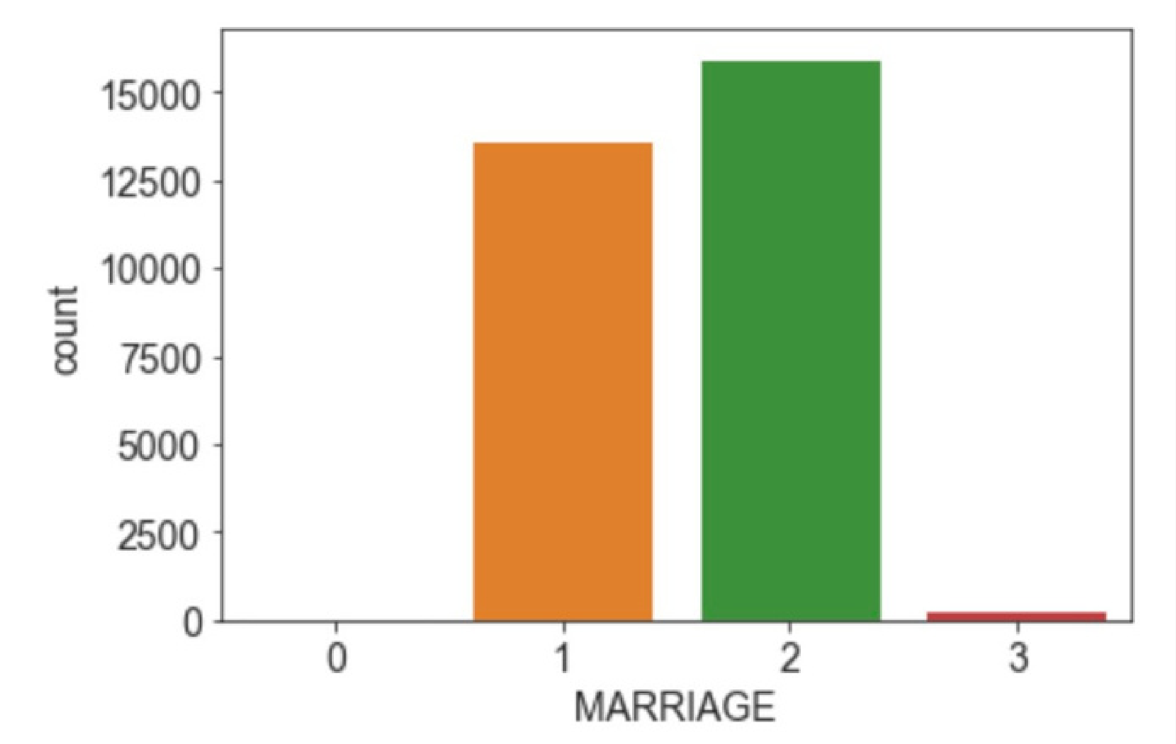
- **Adopt best performing ANN model** with ability to predict 76.21% (AUC) and 81.78% accuracy

- **Improve data quality management** and regularly update customer data

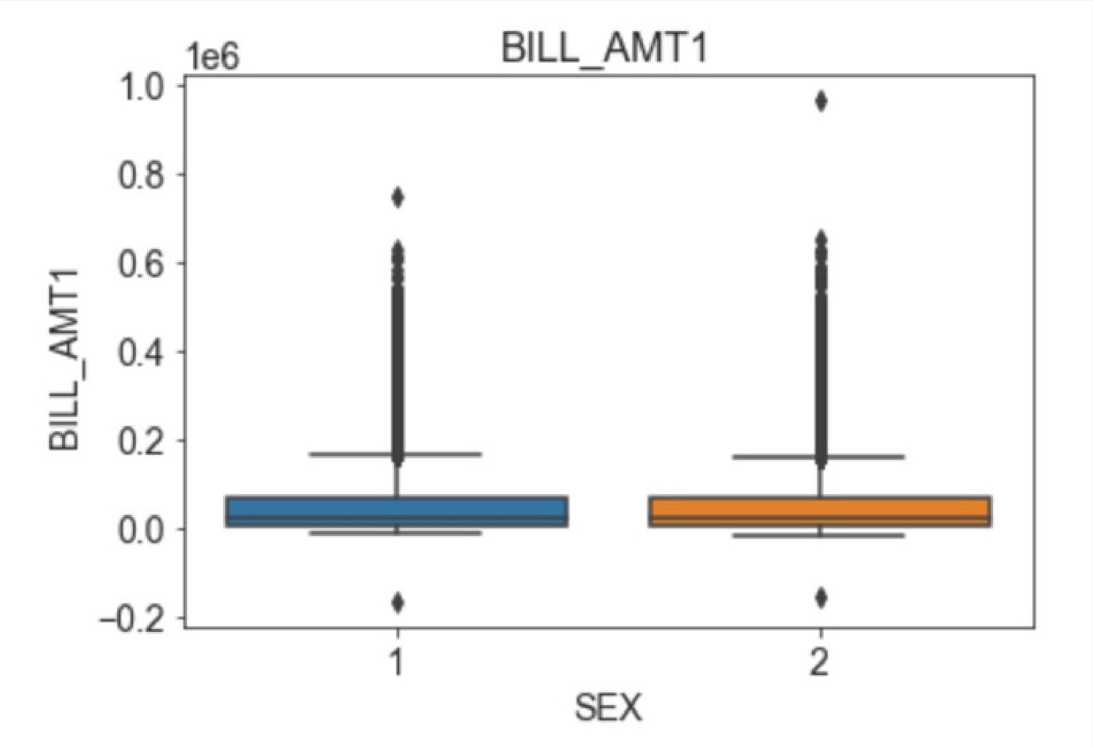
- **Interest should be adjusted** for customer with low credit score and more likely to default
- **Grant small amount loans** rather than big amount.

- **Create creditscore** for each customer and use it to drive decison-making (whether give out loan, the amount safe to give out)

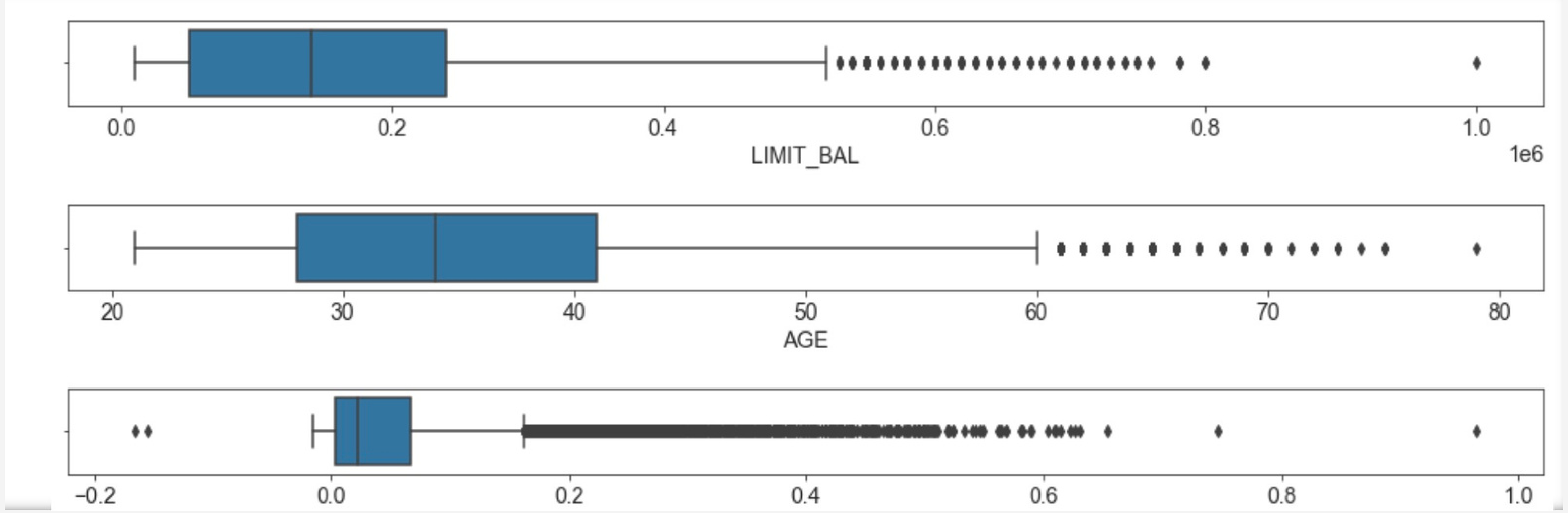
Appendix (1)



Marrital Status Bar Plot

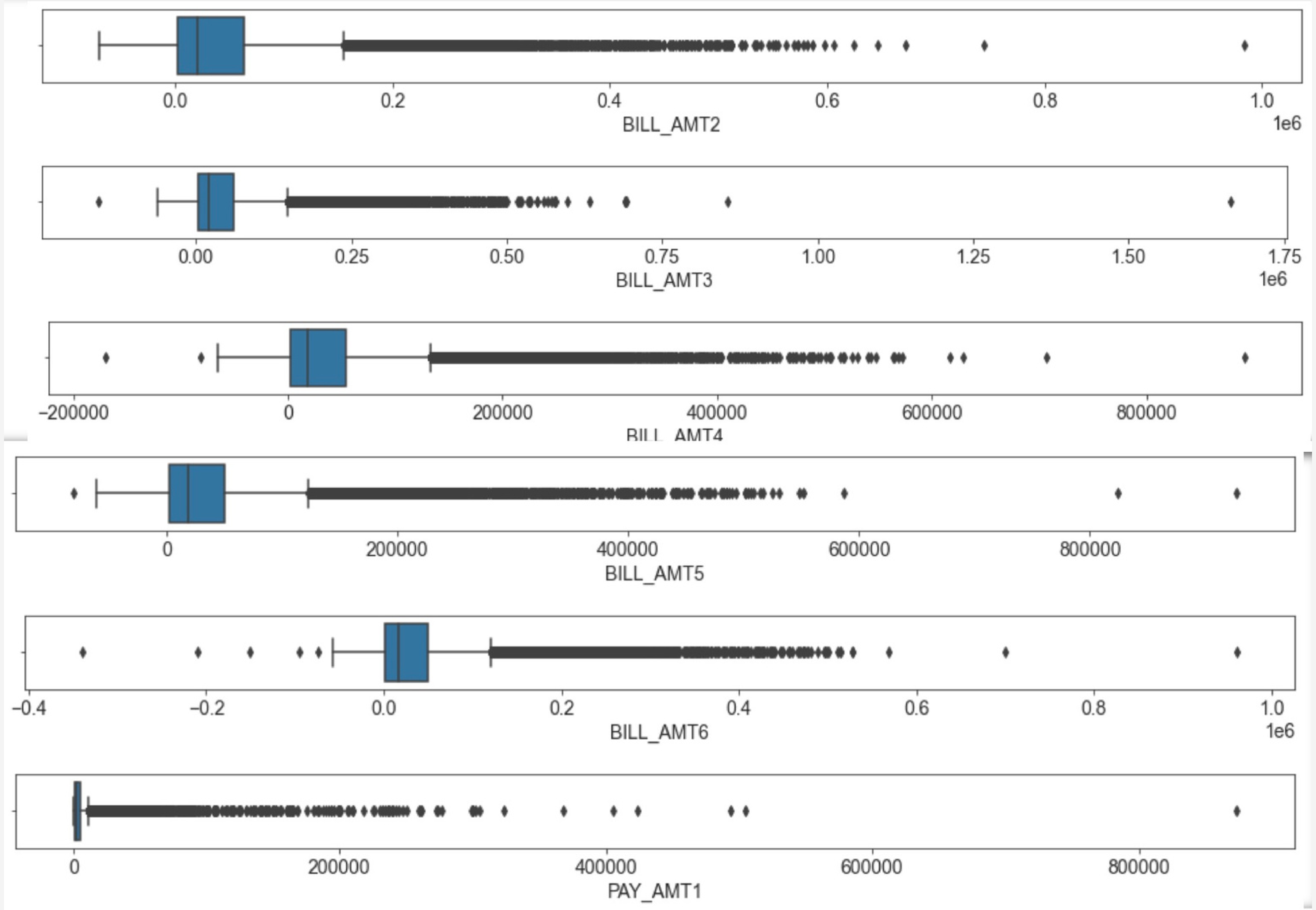


BILL_AMT1 Boxplot by Sex



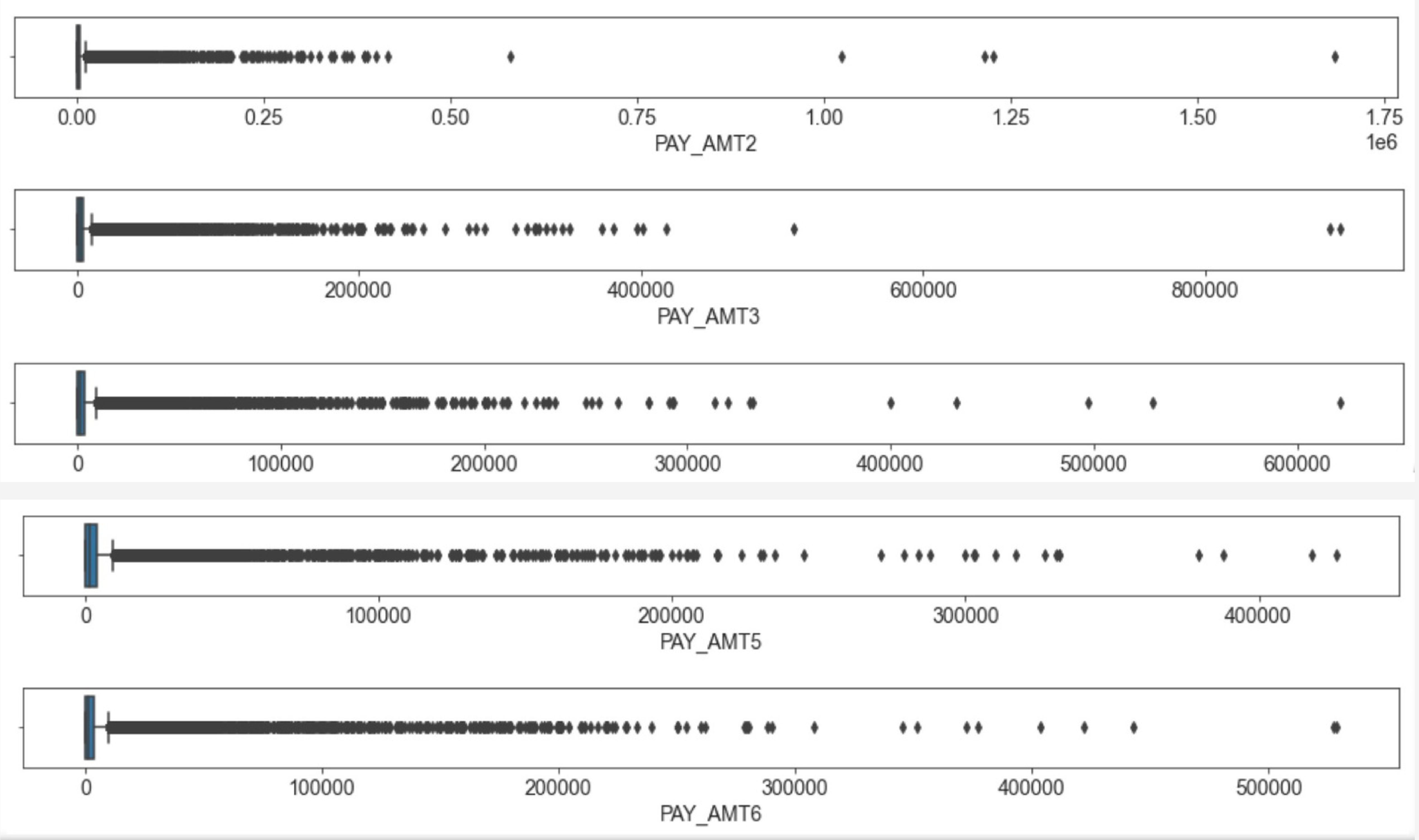
Boxplot of Numeric Variables (1)

Appendix (2)



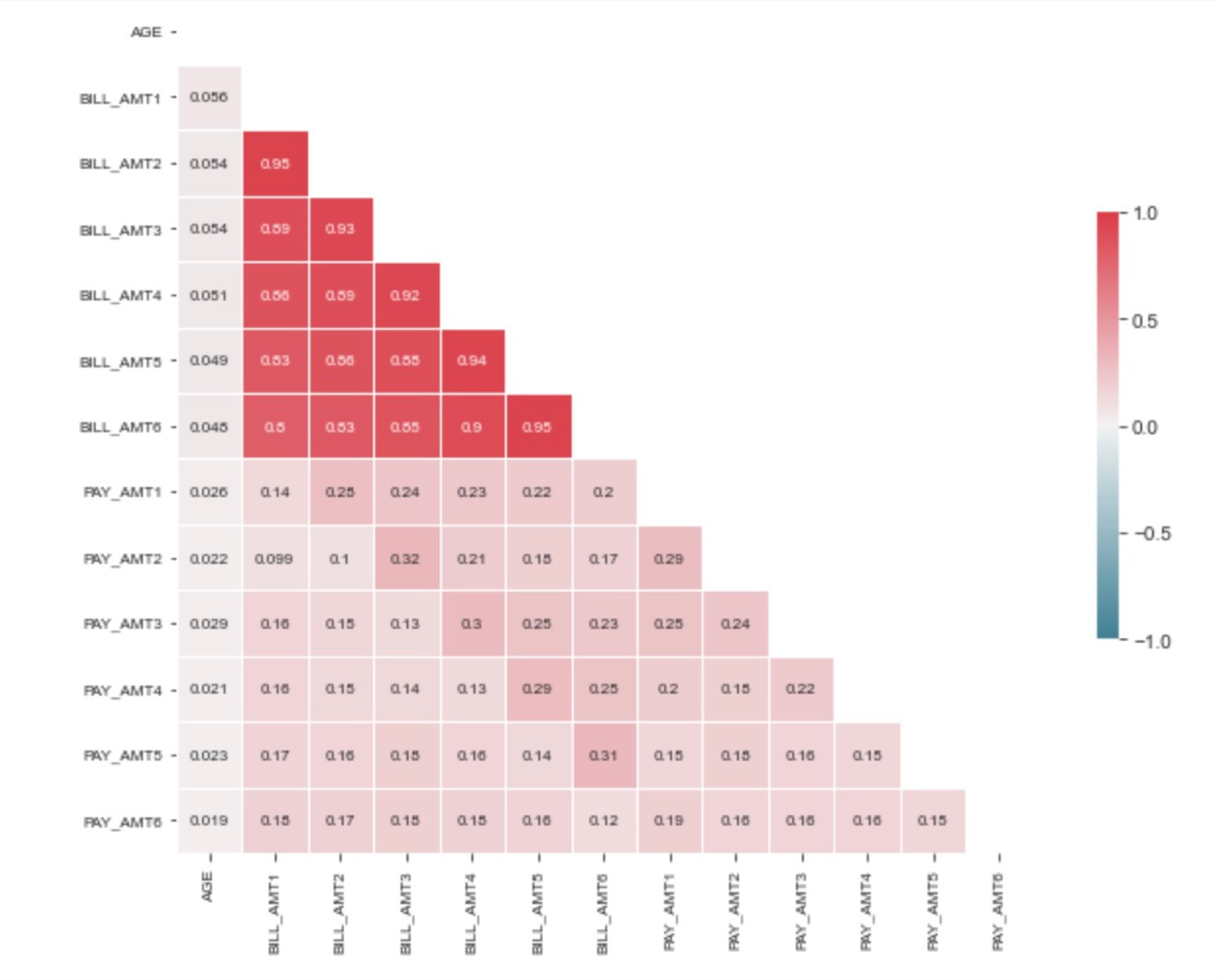
Boxplot of Numeric Variables (2)

Appendix (3)



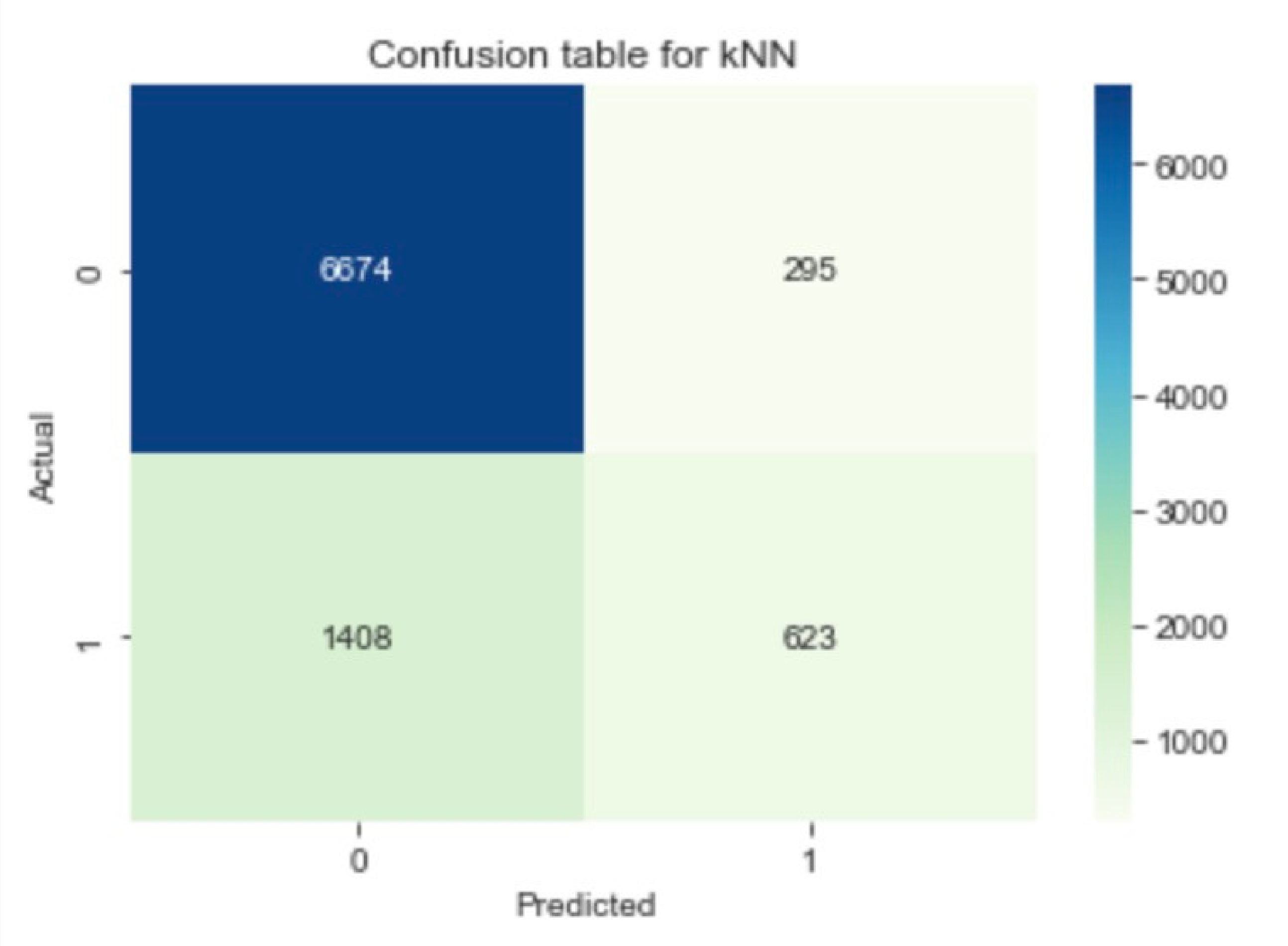
Boxplot of Numeric Variables (3)

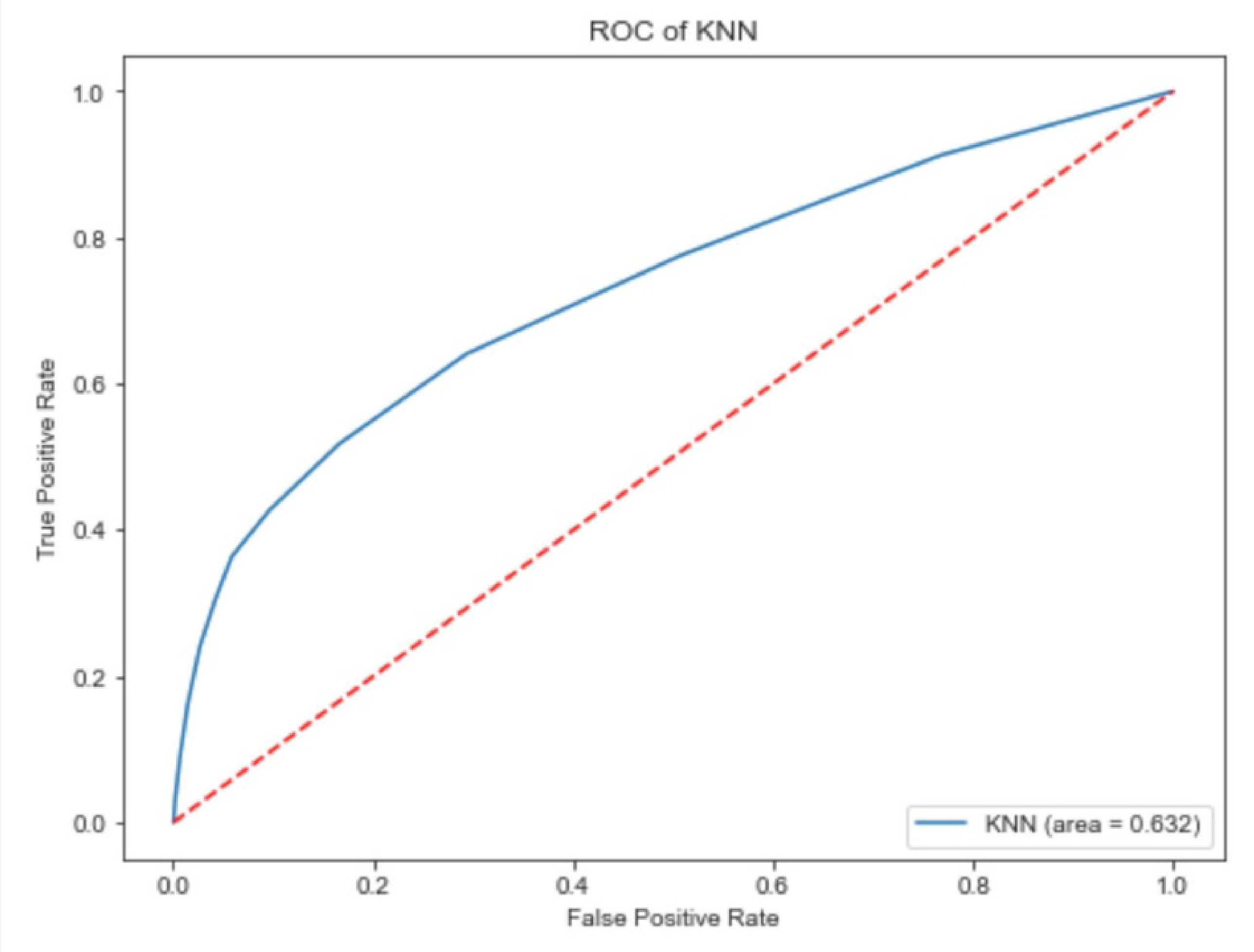
Appendix (4)



Correlation of Numeric Variables

Appendix (5)

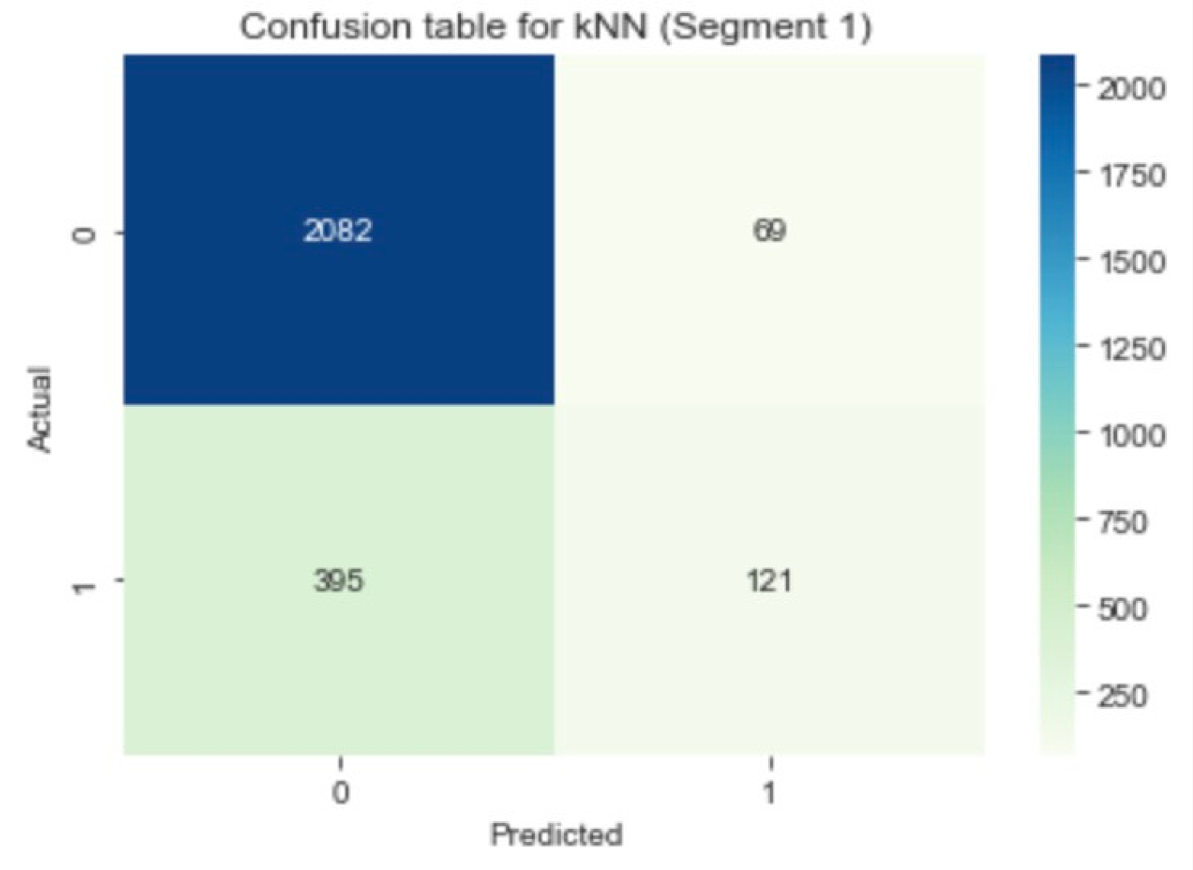


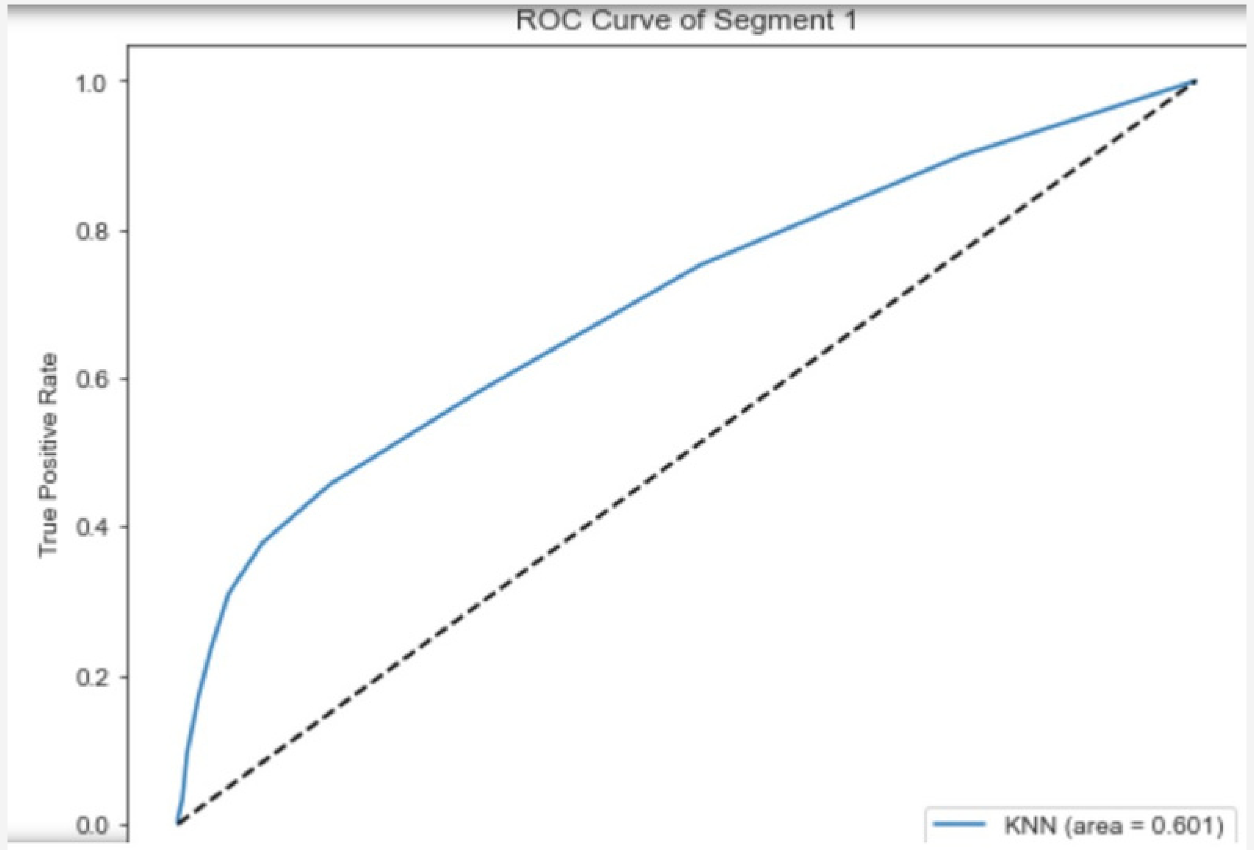



```
Accuracy: 0.8107777777777778
Precision: 0.6786492374727668
Misclassification: 0.18922222222222224
True Positive: 0.3067454455933038
False Positive: 0.04233031998852059
Specificity: 0.9576696800114795
Prevalence: 0.22566666666666665
```

Confusion Table, ROC, Merics of non-segmented KNN Model

Appendix (6)

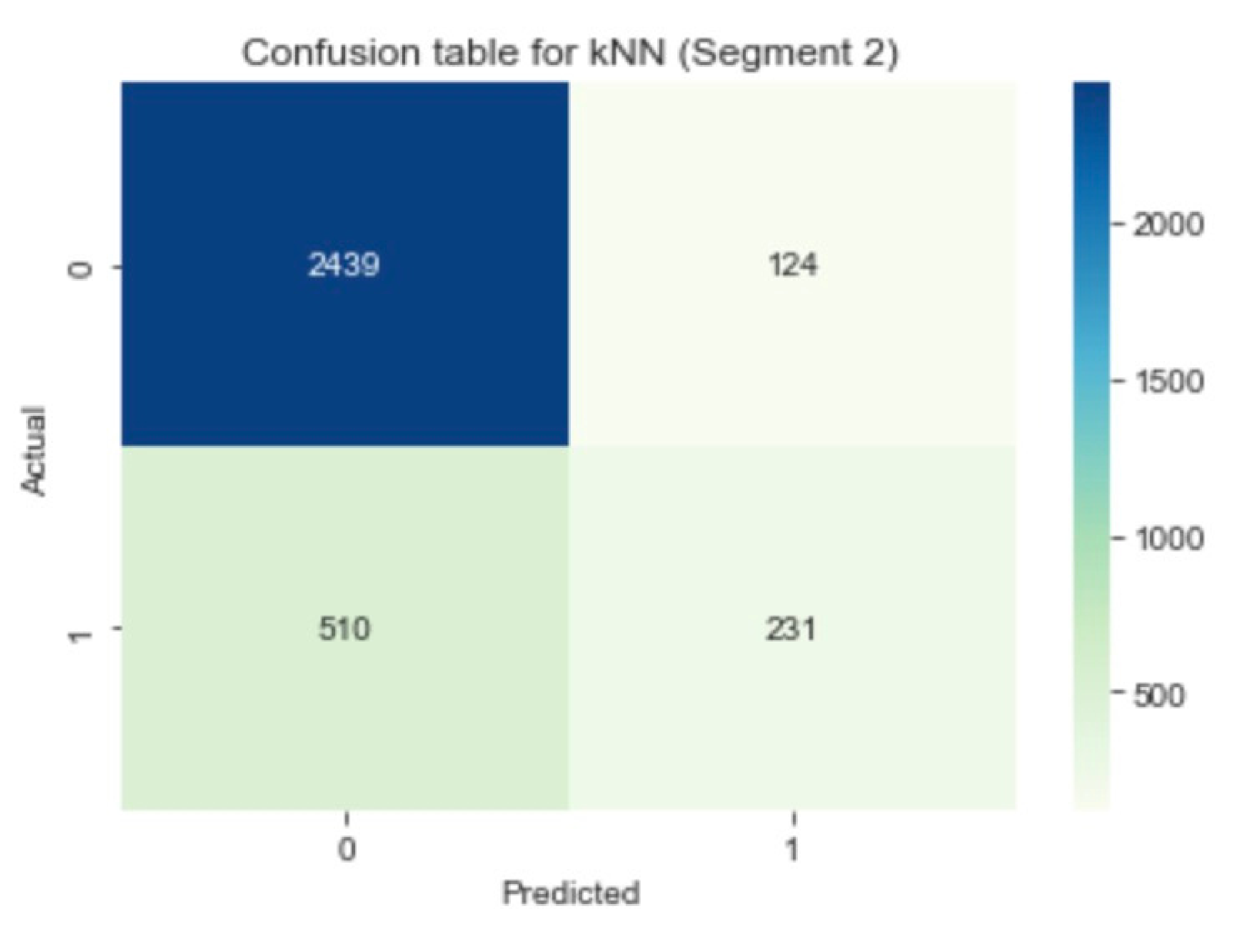


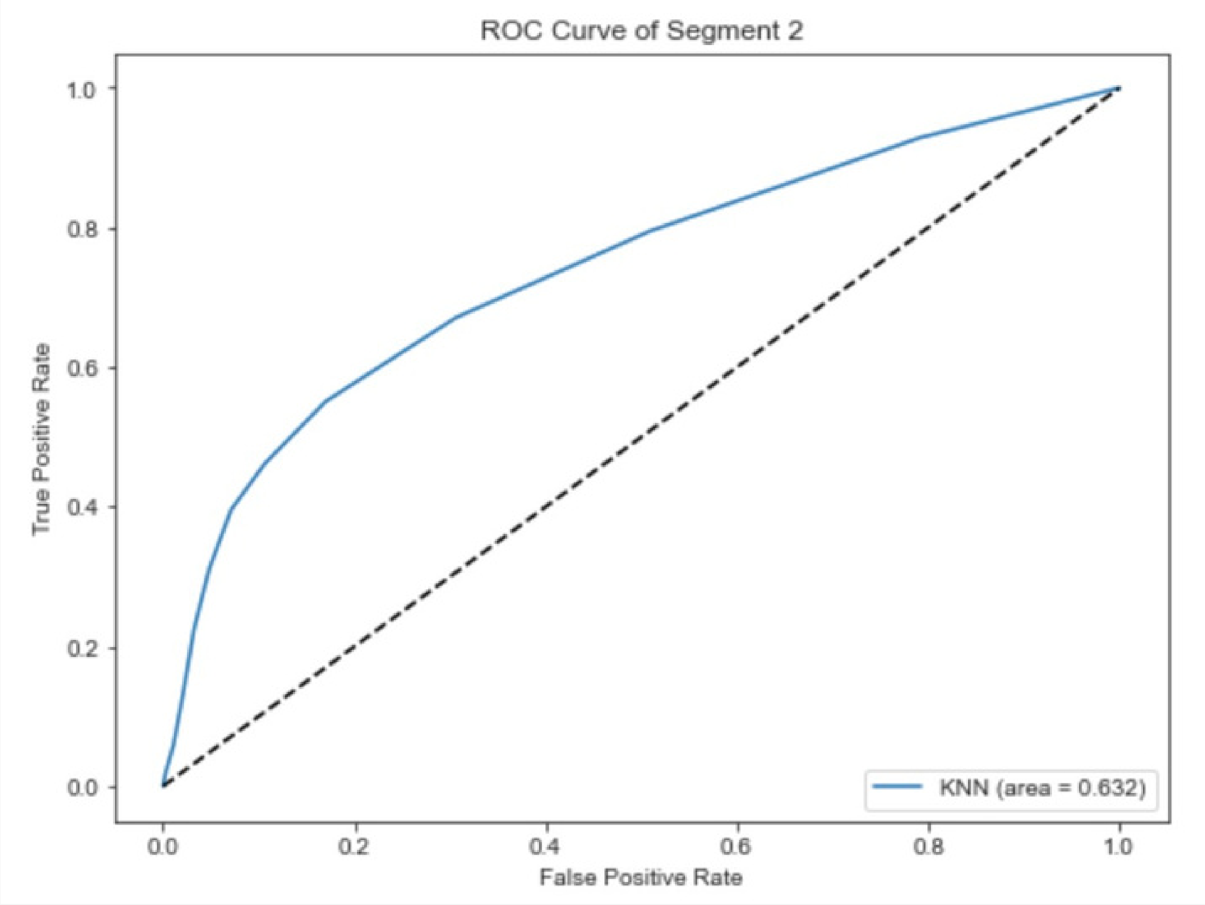



```
Accuracy: 0.8260217472815898
Precision: 0.6368421052631579
Misclassification: 0.1739782527184102
True Positive: 0.23449612403100775
False Positive: 0.03207810320781032
Specificity: 0.9679218967921897
Prevalence: 0.19347581552305962
```

Confusion Table, ROC, Merics of segmented KNN Model (Segment 1)

Appendix (7)

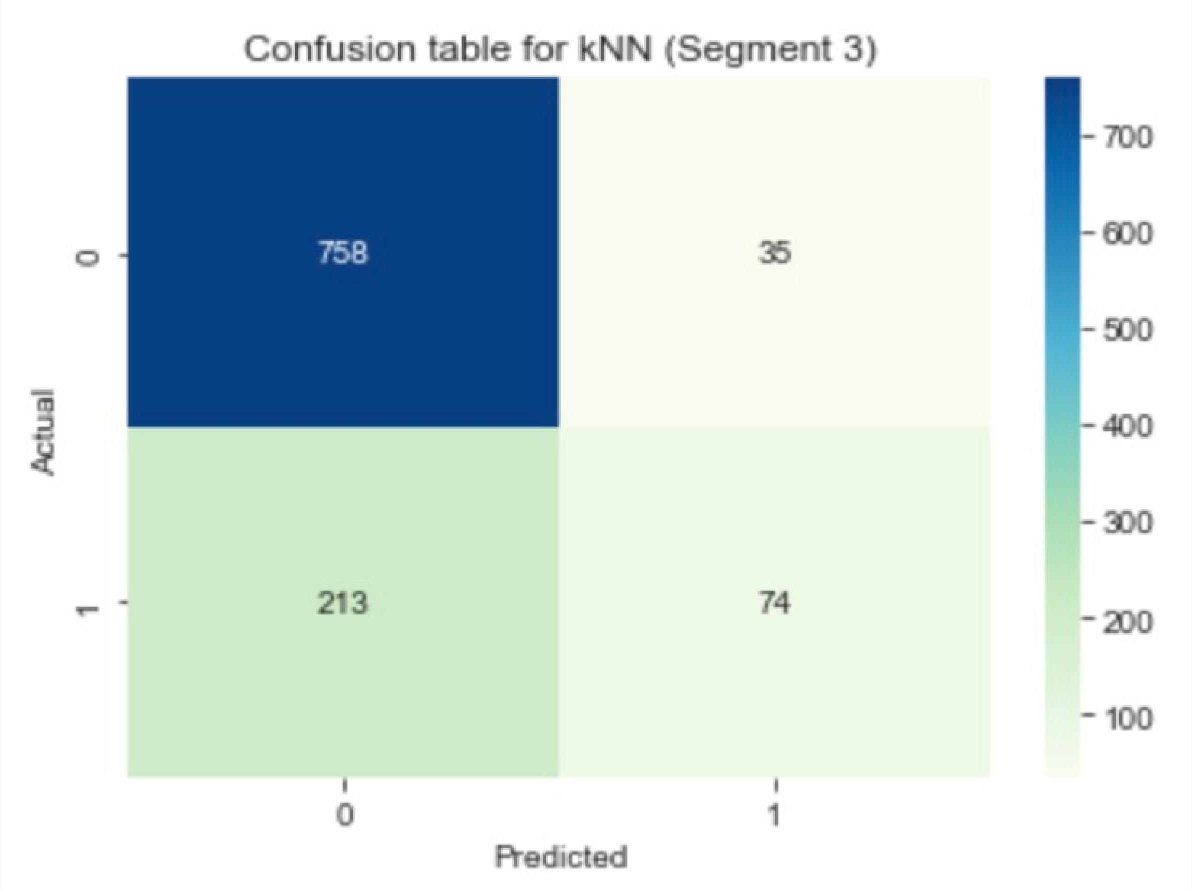


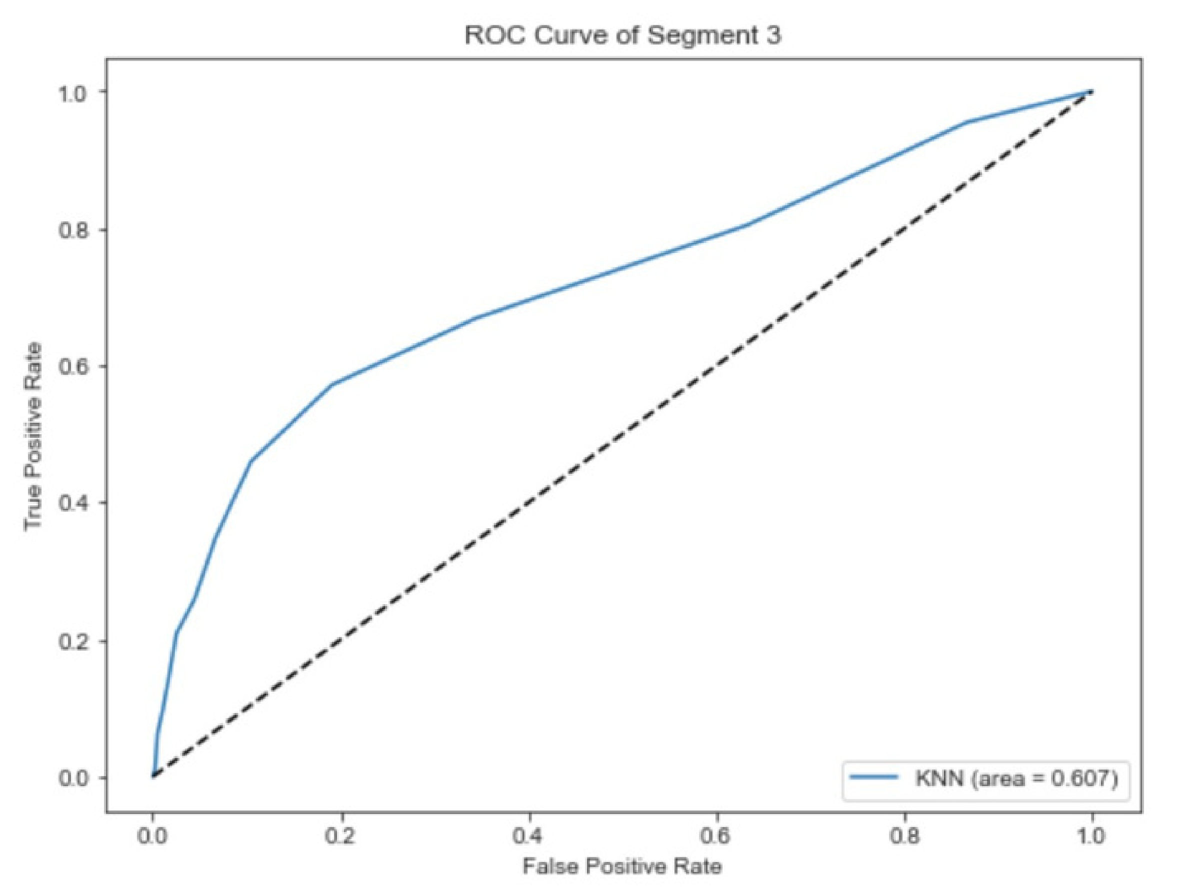



```
Accuracy: 0.8081113801452785
Precision: 0.6507042253521127
Misclassification: 0.19188861985472155
True Positive: 0.3117408906882591
False Positive: 0.048380803745610615
Specificity: 0.9516191962543894
Prevalence: 0.22427360774818403
```

Confusion Table, ROC, Merics of segmented KNN Model (Segment 2)

Appendix (8)

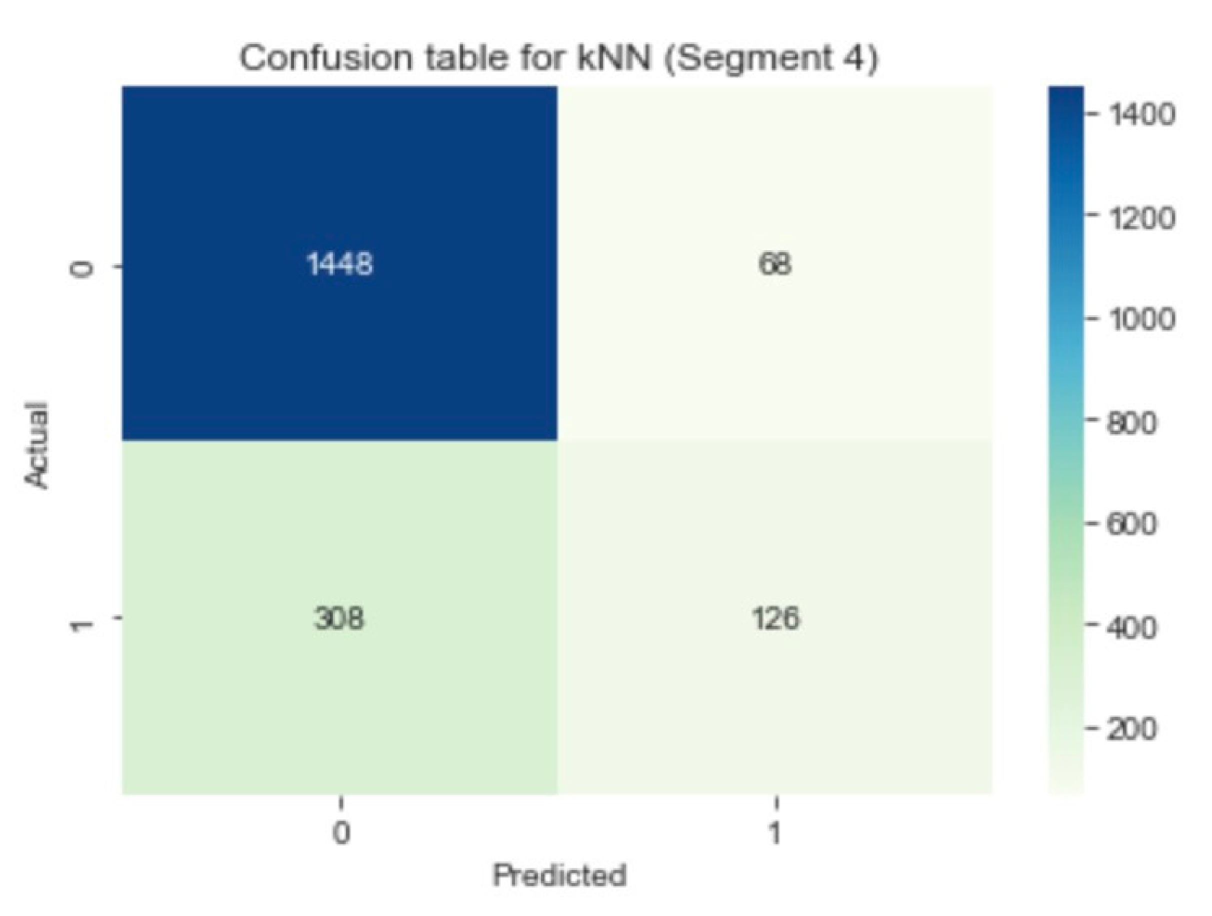


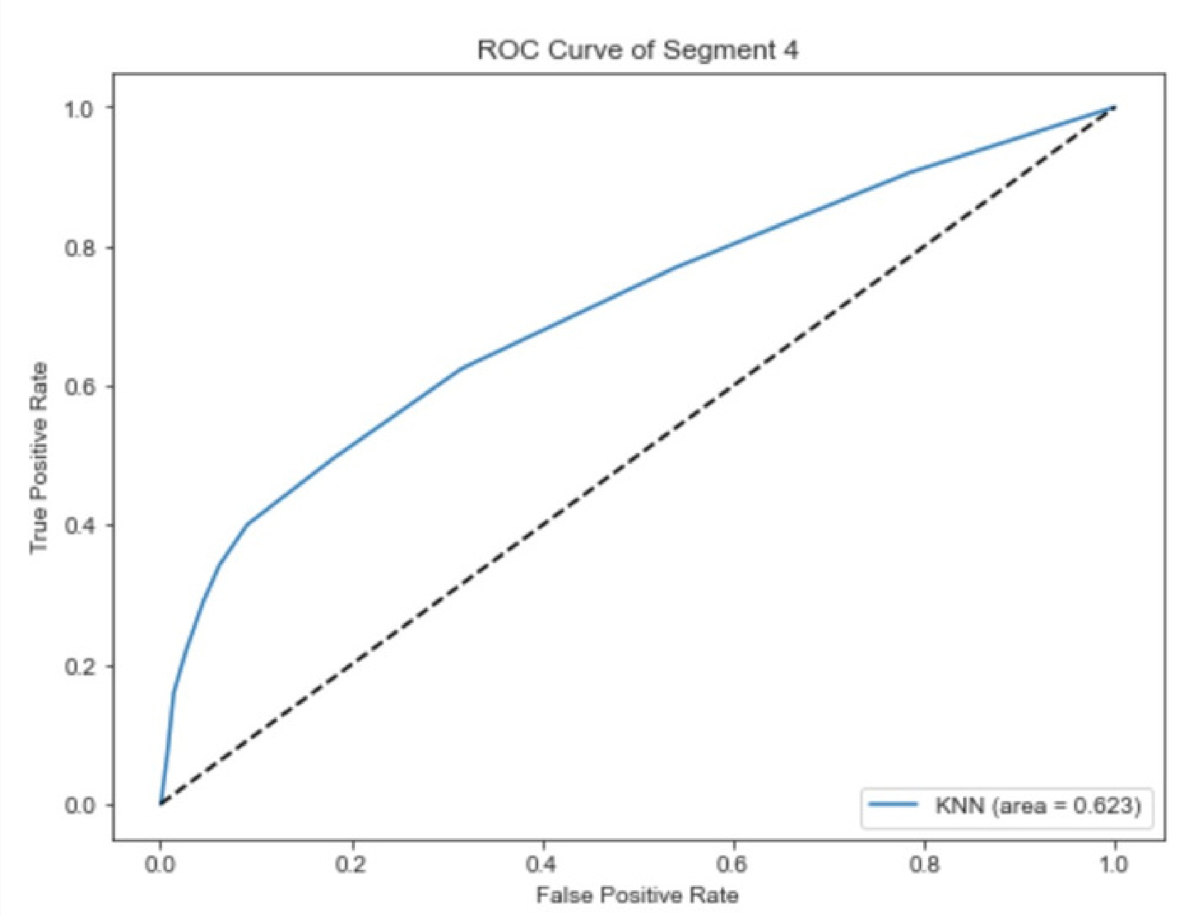



```
Accuracy: 0.7703703703703704
Precision: 0.6788990825688074
Misclassification: 0.22962962962962963
True Positive: 0.2578397212543554
False Positive: 0.044136191677175286
Specificity: 0.9558638083228247
Prevalence: 0.2657407407407407
```

Confusion Table, ROC, Merics of segmented KNN Model (Segment 3)

Appendix (9)

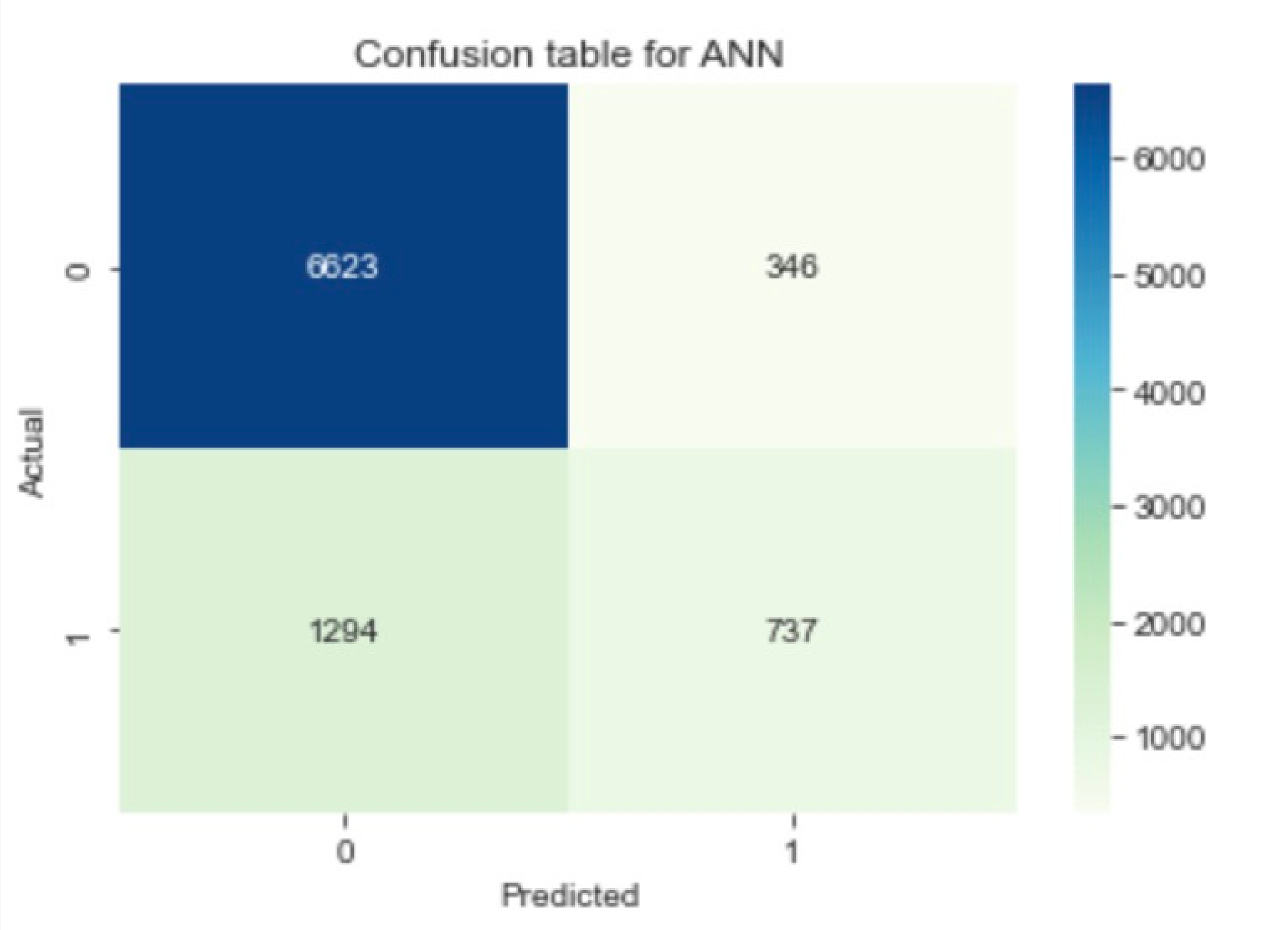


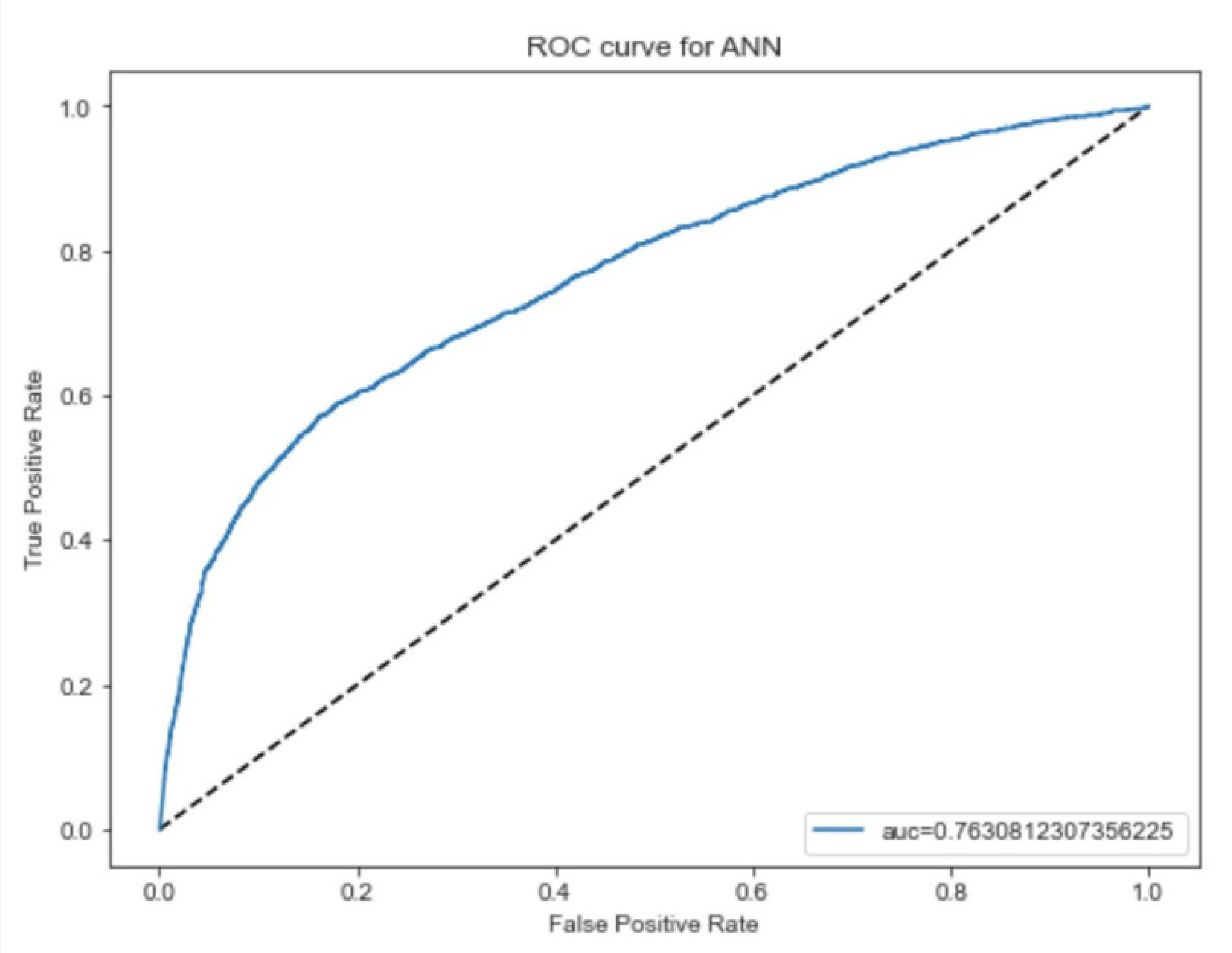



```
Accuracy: 0.8071794871794872
Precision: 0.6494845360824743
Misclassification: 0.19282051282051282
True Positive: 0.2903225806451613
False Positive: 0.044854881266490766
Specificity: 0.9551451187335093
Prevalence: 0.22256410256410256
```

Confusion Table, ROC, Merics of segmented KNN Model (Segment 4)

Appendix (10)






```
Accuracy: 0.8177777777777778
Missclassification Rate: 0.18222222222222223
True Positive Rate: 0.8365542503473538
False Positive Rate: 0.31948291782086796
Specificity: 0.6805170821791321
Precision: 0.9503515568948199
Prevalence: 0.7743333333333333
```

Confusion Table, ROC, Merics of ANN Model

Appendix (11)

	Segment 1	Segment 2	Segment 3	Segment 4	KNN	ANN
Accuracy	0.826022	0.808111	0.770370	0.807179	0.810778	0.817778
Misclassification	0.173978	0.191889	0.229630	0.192821	0.189222	0.182222
True Positive	0.234496	0.311741	0.257840	0.290323	0.306745	0.836554
False Positive	0.032078	0.048381	0.044136	0.044855	0.042330	0.319483
Specificity	0.967922	0.951619	0.955864	0.955145	0.957670	0.680517
Precision	0.636842	0.650704	0.678899	0.649485	0.678649	0.950352
Prevalence	0.193476	0.224274	0.265741	0.222564	0.225667	0.774333
AUC	0.601209	0.631680	0.606852	0.622734	0.632208	0.763081

Model Merics Comparison