Default Payments of Credit Cards

BY – BHAVYA BALASUBRAMANYA

JAN 2019

Supervised Learning Capstone

This capstone was a part of supervised learning curriculum at Thinkful.

For the entire project please visit my repository

Outline

- Introduction
- Understanding the dataset
- Exploratory Data Analysis
- Predicting Default payments
- Feature engineering
- Iteration and evaluation of data
- Best solution for our predictions
- Conclusion
- Next steps

Introduction

What does it mean to default on a credit card?

In every country there is a specific time period after which if a person fails to make any payment, the lender assumes that the person is never going to pay and moves the status of the loan as defaulted. At this point, the lender will typically close the account, write off the debt as bad debt and sell the account to collection agency. If there is a continued non payment, the credit scores are negatively affected in different credit bureaus.

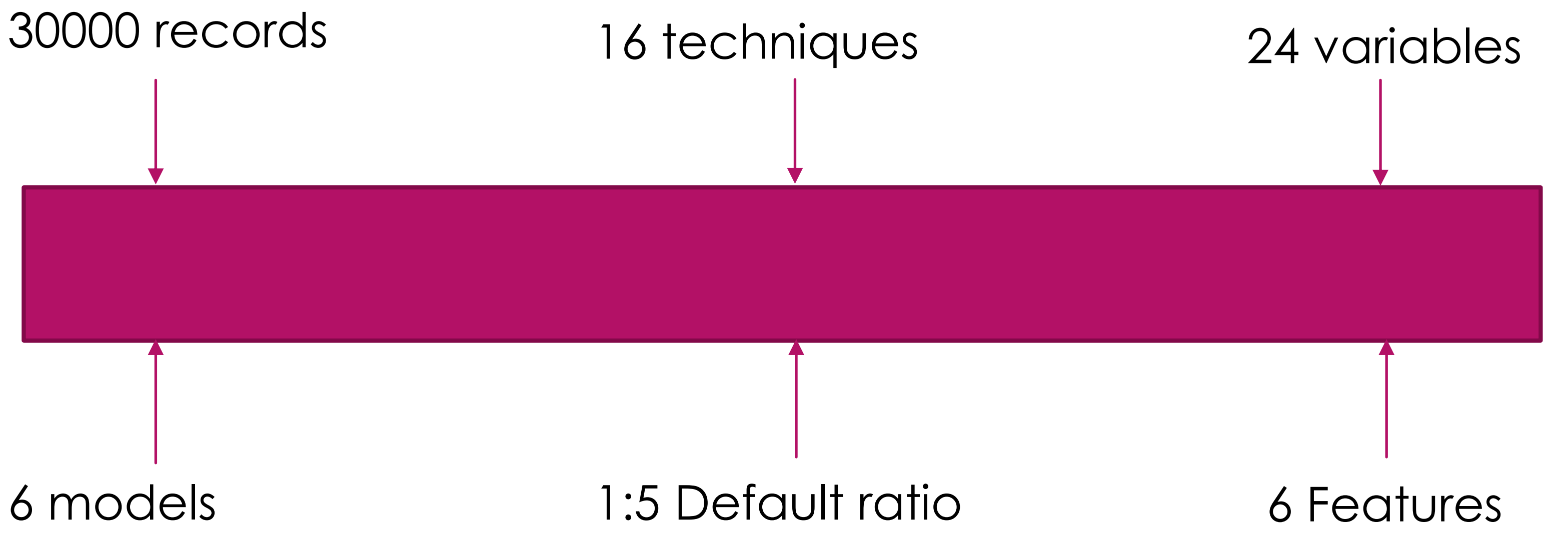
Understanding the dataset

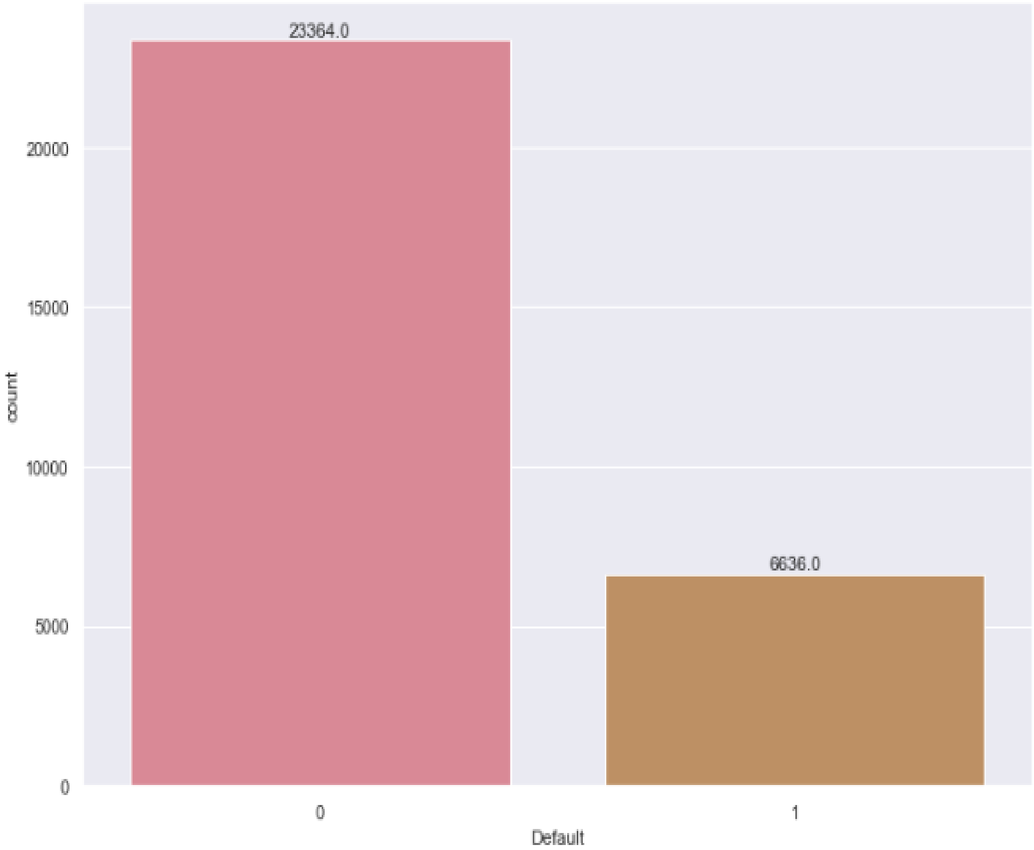
Our dataset consists of information on **default payments, demographic factors, credit data, history of payment, bill statements** of credit card clients in **Taiwan** from **April 2005** to **September 2005**.

We can divide our data attributes into 5 segments:

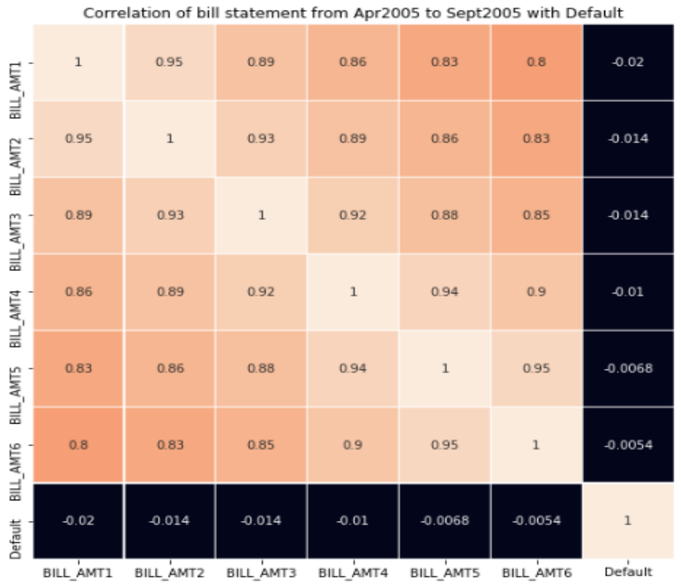
- Bill amount data for above months.
- Payment amount for the said months that was done.
- (Re)Payment Indicator for the above 6 months.
- Demographic data like age, sex, credit limit, marital status and education.
- Default indicator which is our target.

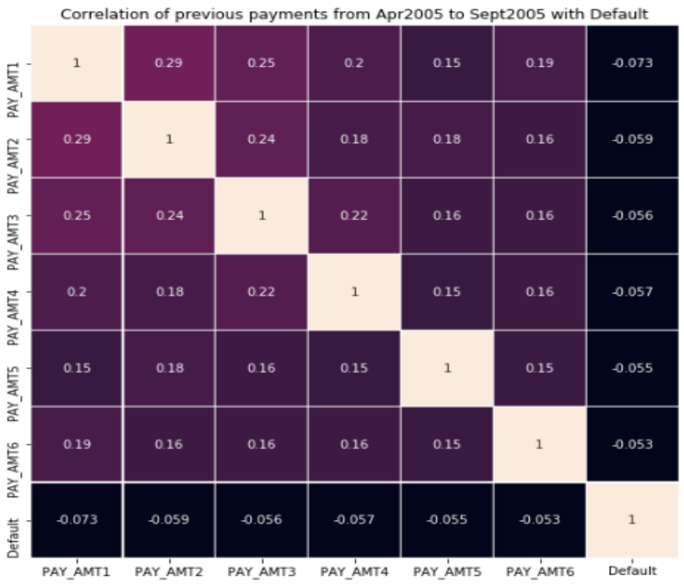
Exploratory Data Analysis



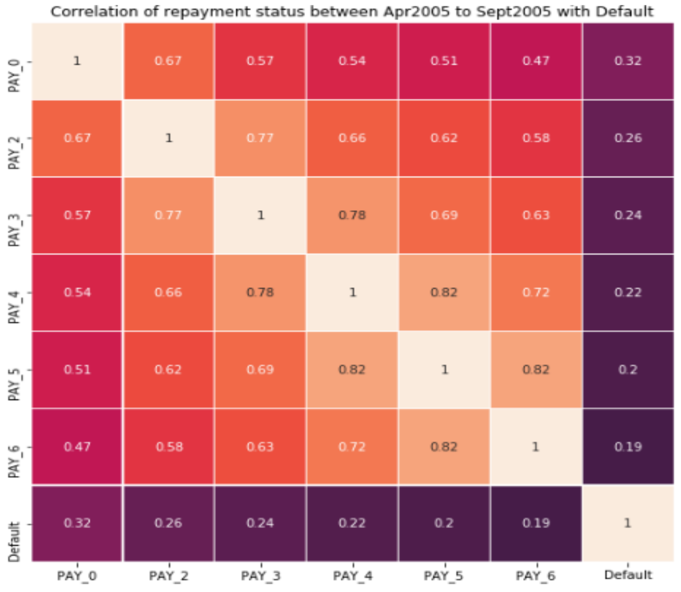


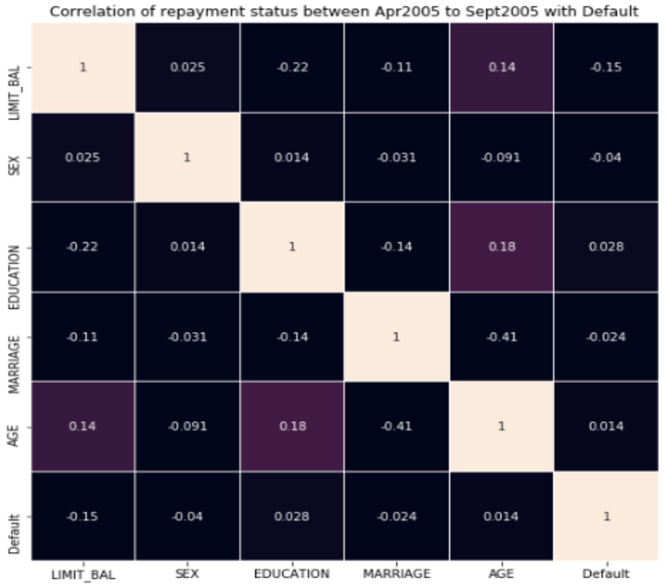
How Does data look?


Correlation of bill statement from Apr2005 to Sept2005 with Default
BILL_AMT1 1 0.95 0.89 0.86 0.83 0.8 -0.02
BILL_AMT2 0.95 1 0.93 0.89 0.86 0.83 -0.014
BILL_AMT3 0.89 0.93 1 0.92 0.88 0.85 -0.014
BILL_AMT4 0.86 0.89 0.92 1 0.94 0.9 -0.01
BILL_AMT5 0.83 0.86 0.88 0.94 1 0.95 -0.0068
BILL_AMT6 0.8 0.83 0.85 0.9 0.95 1 -0.0054
Default -0.02 -0.014 -0.014 -0.01 -0.0068 -0.0054 1
BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6 Default

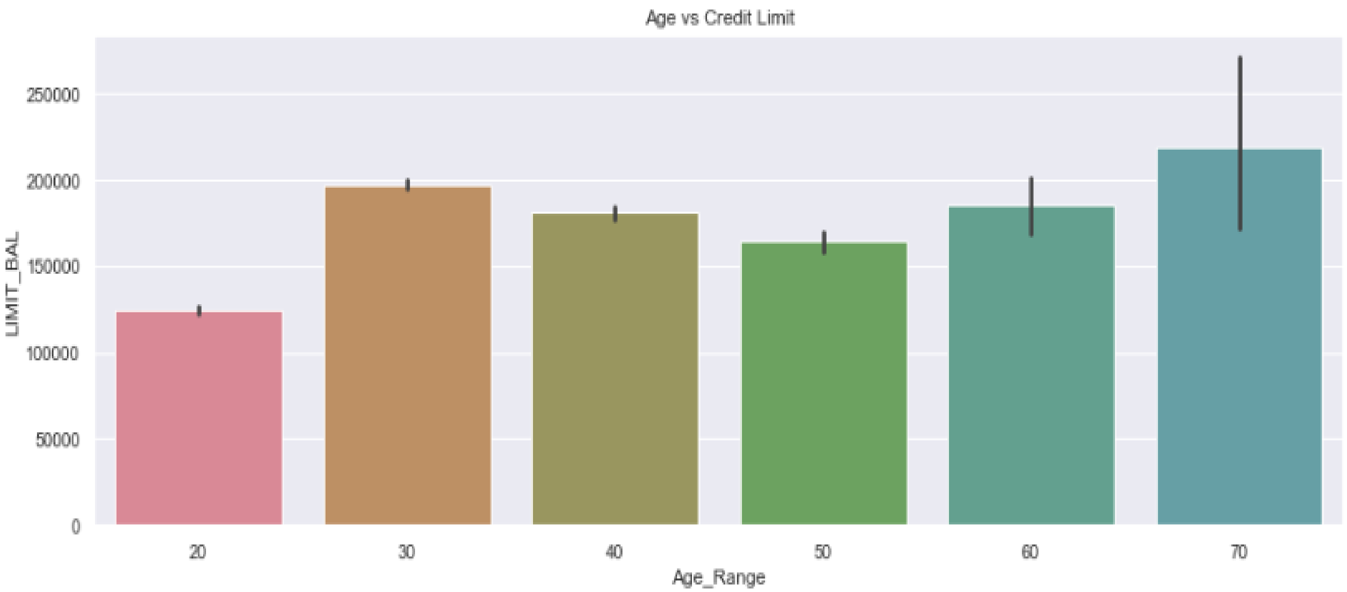

Correlation of previous payments from Apr2005 to Sept2005 with Default
PAY_AMT1 1 0.29 0.25 0.2 0.15 0.19 -0.073
PAY_AMT2 0.29 1 0.24 0.18 0.18 0.16 -0.059
PAY_AMT3 0.25 0.24 1 0.22 0.16 0.16 -0.056
PAY_AMT4 0.2 0.18 0.22 1 0.15 0.16 -0.057
PAY_AMT5 0.15 0.18 0.16 0.15 1 0.15 -0.055
PAY_AMT6 0.19 0.16 0.16 0.16 0.15 1 -0.053
Default -0.073 -0.059 -0.056 -0.057 -0.055 -0.053 1
PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6 Default

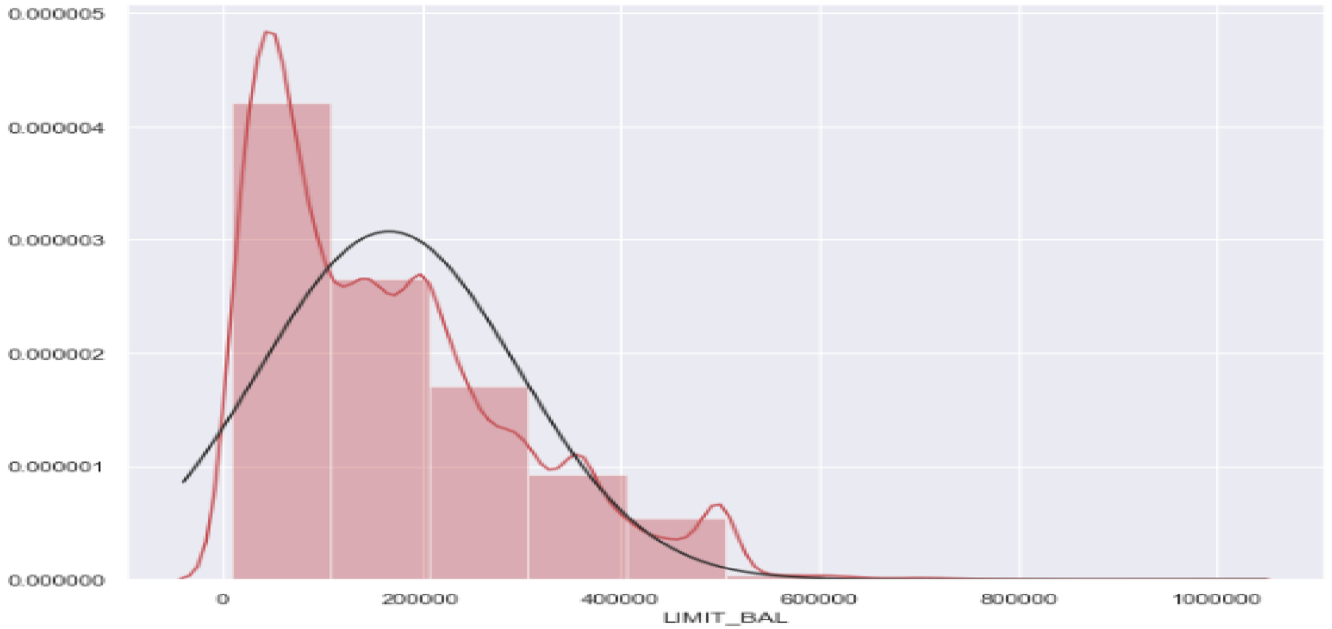
Correlated?

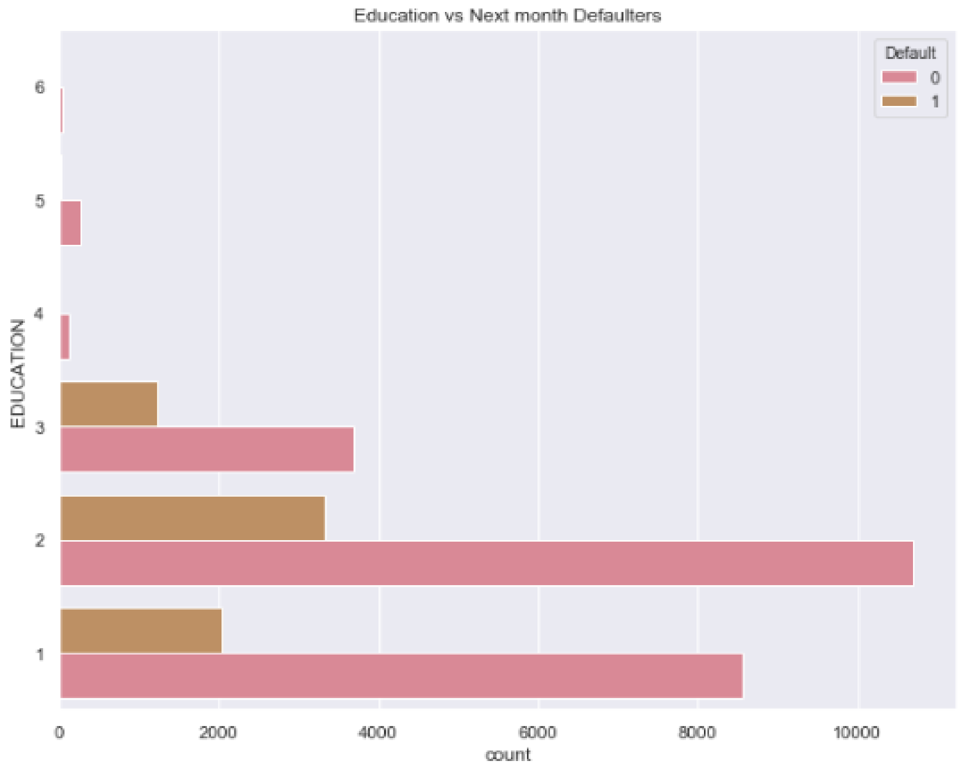

Correlation of repayment status between Apr2005 to Sept2005 with Default
PAY_0 1 0.67 0.57 0.54 0.51 0.47 0.32
PAY_2 0.67 1 0.77 0.66 0.62 0.58 0.26
PAY_3 0.57 0.77 1 0.78 0.69 0.63 0.24
PAY_4 0.54 0.66 0.78 1 0.82 0.72 0.22
PAY_5 0.51 0.62 0.69 0.82 1 0.82 0.2
PAY_6 0.47 0.58 0.63 0.72 0.82 1 0.19
Default 0.32 0.26 0.24 0.22 0.2 0.19 1
PAY_0 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6 Default

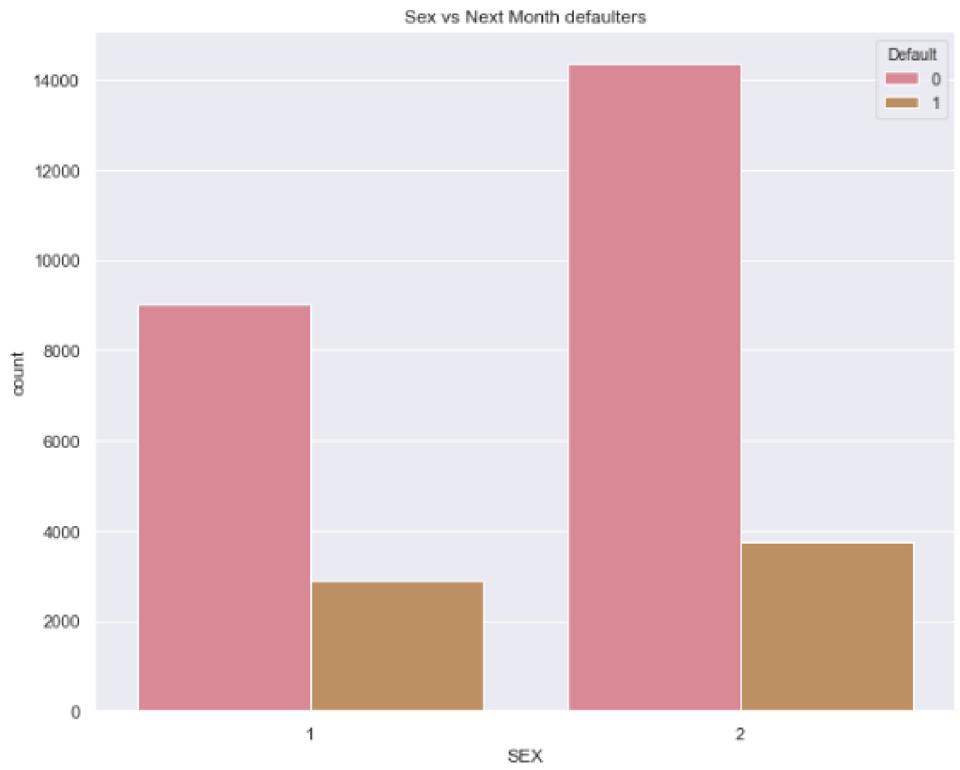

Correlation of repayment status between Apr2005 to Sept2005 with Default
LIMIT_BAL 1 0.025 -0.22 -0.11 0.14 -0.15
SEX 0.025 1 0.014 -0.031 -0.091 -0.04
EDUCATION -0.22 0.014 1 -0.14 0.18 0.028
MARRIAGE -0.11 -0.031 -0.14 1 -0.41 -0.024
AGE 0.14 -0.091 0.18 -0.41 1 0.014
Default -0.15 -0.04 0.028 -0.024 0.014 1
LIMIT_BAL SEX EDUCATION MARRIAGE AGE Default

- Which age group gets more credit limit?
- What is the typical credit limit?

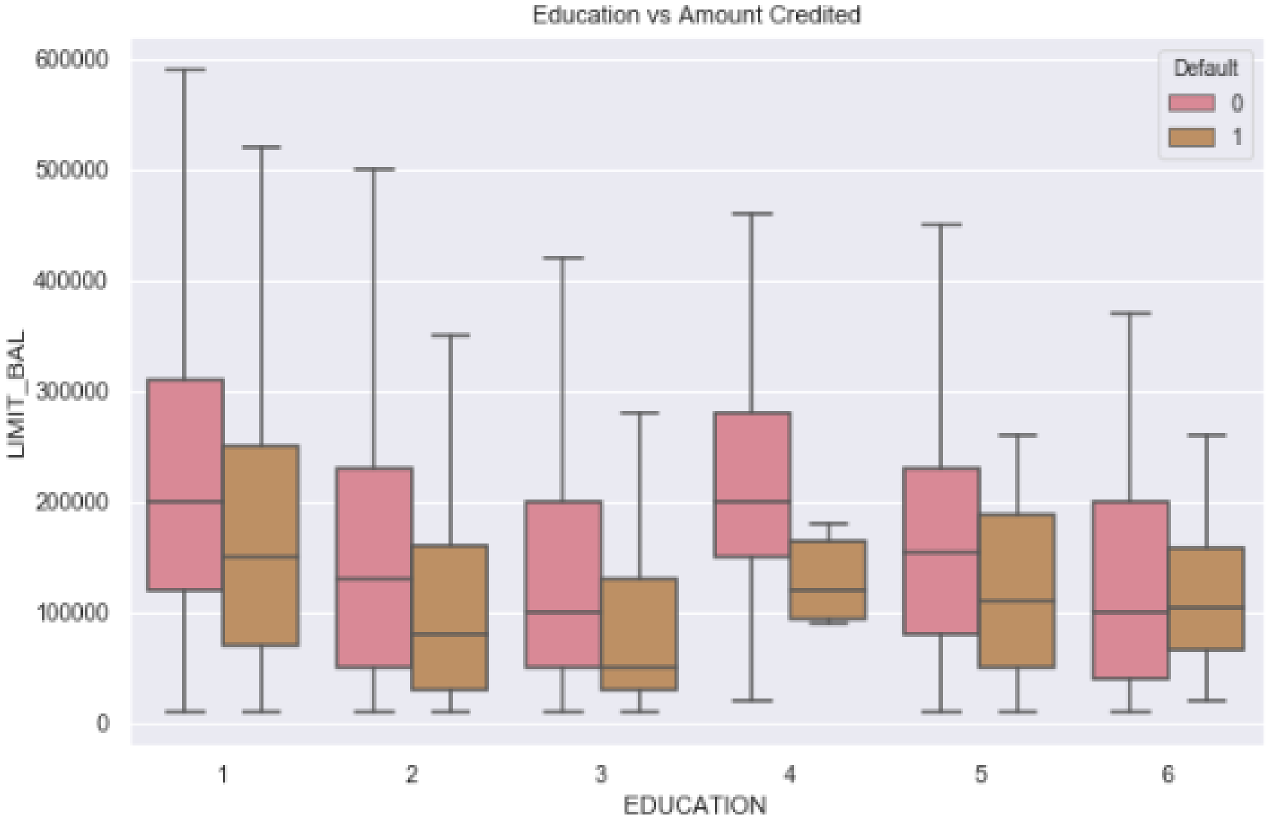




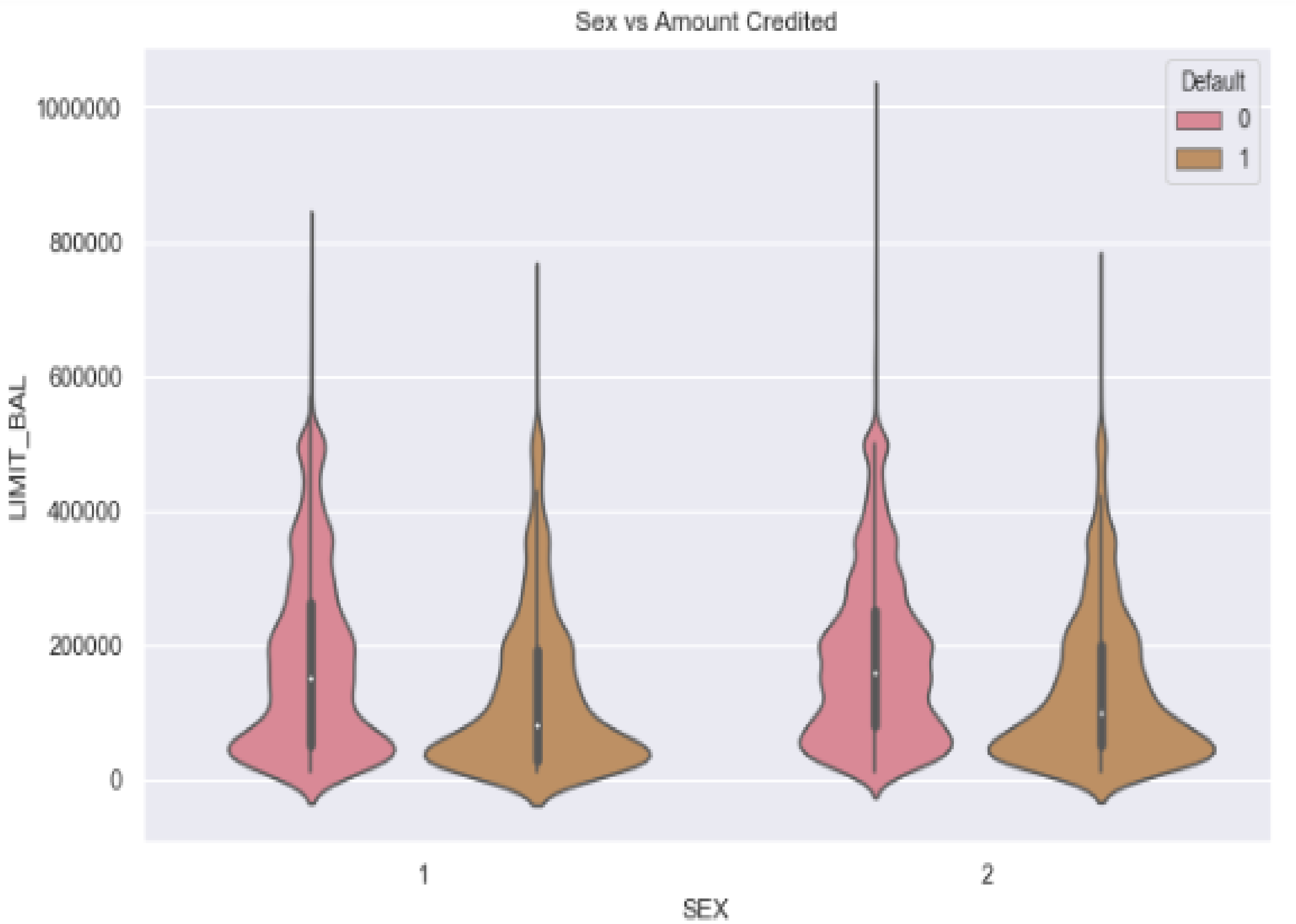




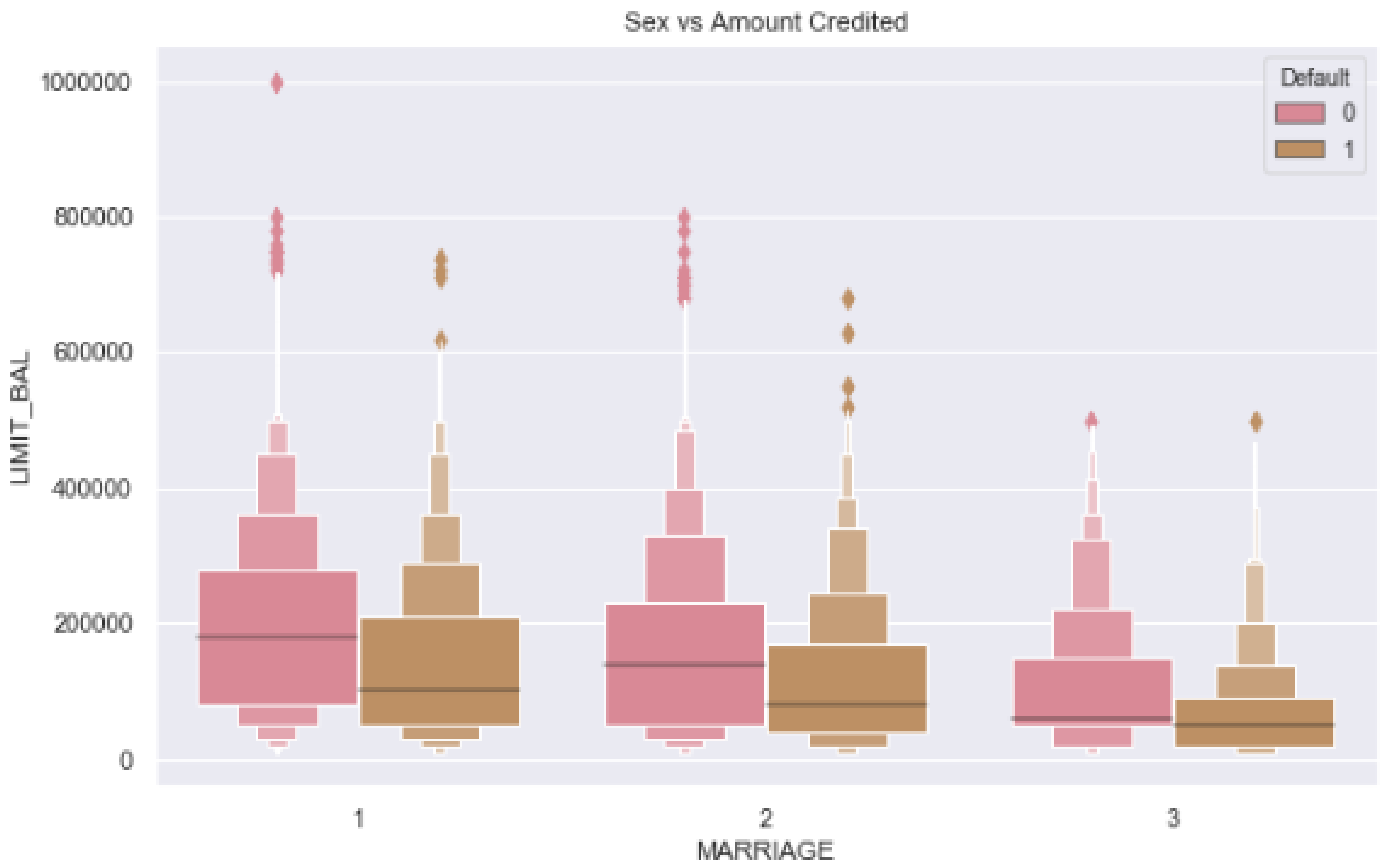
- How does education affect Defaulting?
- Who are more likely to default Women or Men?



What education gets more credit limit?



Which sex gets more credit limit?

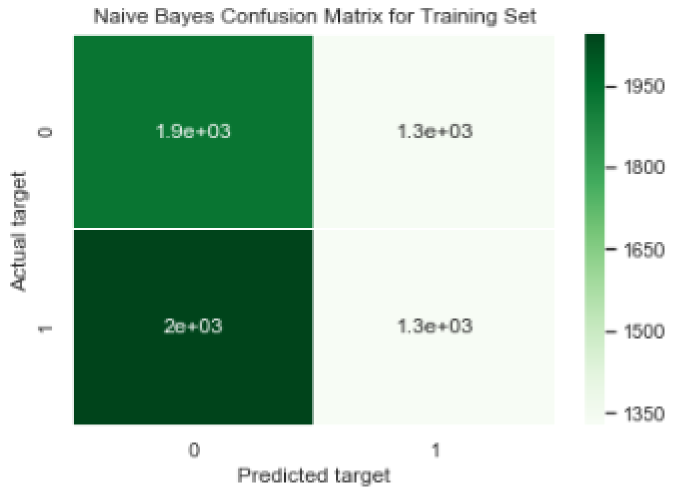


What marital status gets more credit limit?

Predicting Default Payments

Naïve Bayes Model

- ✓ With training and test data: 1:1 ratio.
- ✓ Bernoulli's method.
- ✓ Average accuracy = 67.8%
- ✓ Root mean squared error = 71.3%

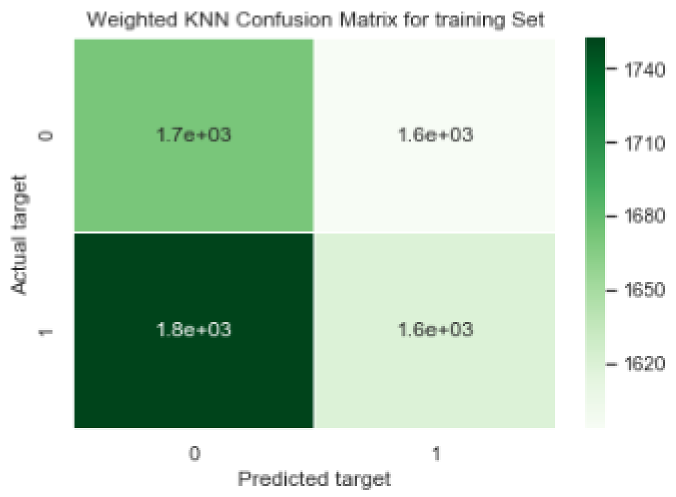


Predicting Default Payments Cont..

KNN Model

i. Weighted

- ✓ With scaling and PCA on entire data except target.
- ✓ With training and test data sets on PCA - 1:1 ratio.
- ✓ Average accuracy = 8.2%
- ✓ Root mean squared error = 59.8%

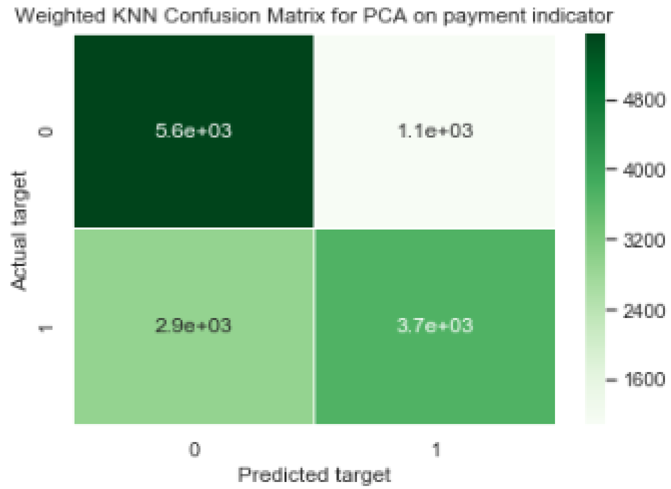


Predicting Default Payments Cont..

KNN Model

i. Weighted

- ✓ With PCA on Payment Indicator.
- ✓ Average accuracy = 0.7%
- ✓ Root mean squared error = 47.1%

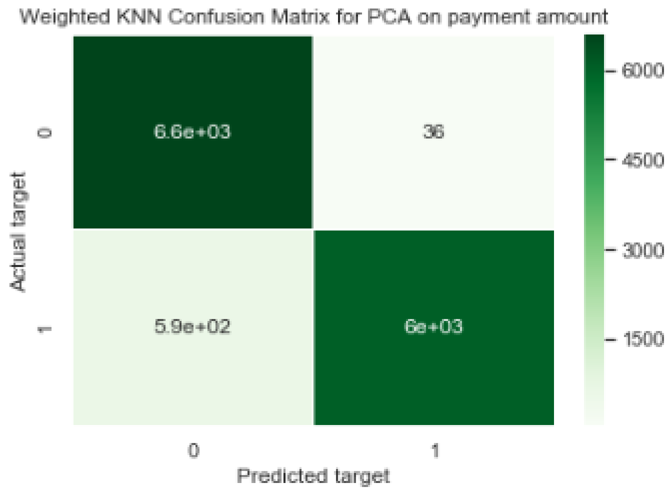


Predicting Default Payments Cont..

KNN Model

i. Weighted

- ✓ With PCA on Payment Amount.
- ✓ Average accuracy = -4.5%
- ✓ Root mean squared error = 14.6%

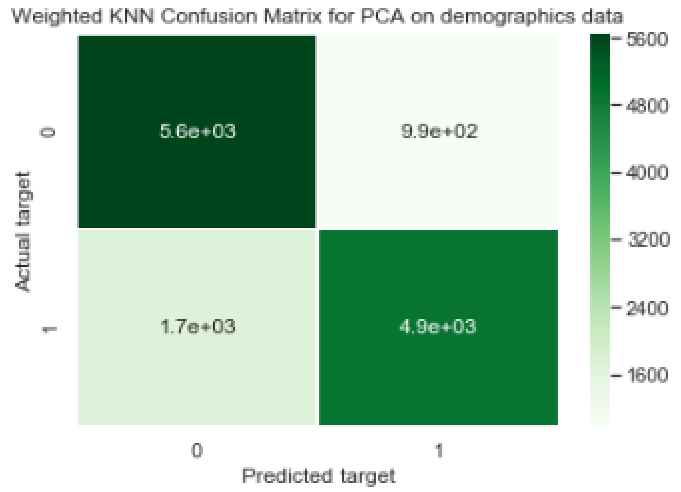


Predicting Default Payments Cont..

KNN Model

i. Weighted

- ✓ With PCA on Demographic data.
- ✓ Average accuracy = -7.8%
- ✓ Root mean squared error = 34.9%

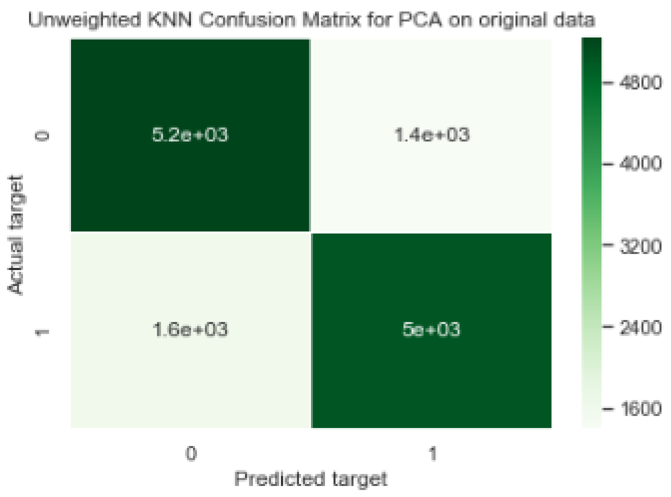


Predicting Default Payments Cont..

KNN Model

ii. Unweighted

- ✓ With PCA on original data.
- ✓ Average accuracy = 1.7%
- ✓ Root mean squared error = 38.8%



Predicting Default Payments Cont..

SVC Model

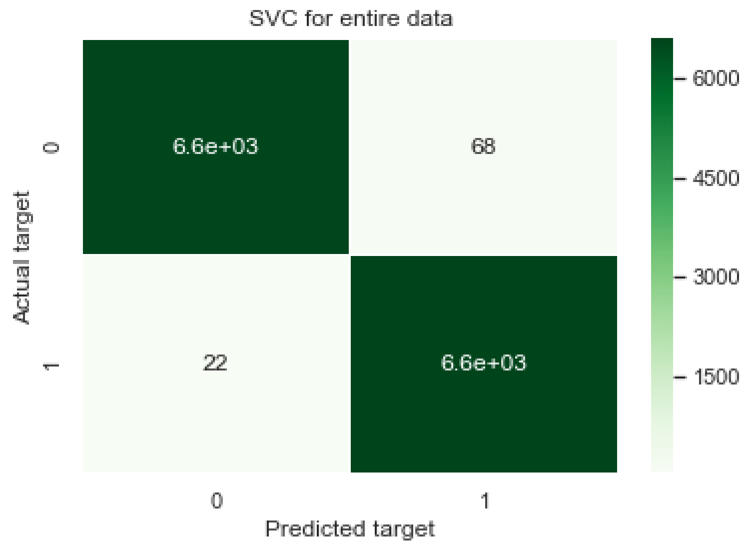
- ✓ With original data.
- ✓ Average accuracy = 51.5%
- ✓ Root mean squared error = 7.1%

Predicting Default Payments Cont..

SVC Model

i. With original data.

- ✓ Average accuracy = 51.3%
- ✓ Root mean squared error = 8.2%

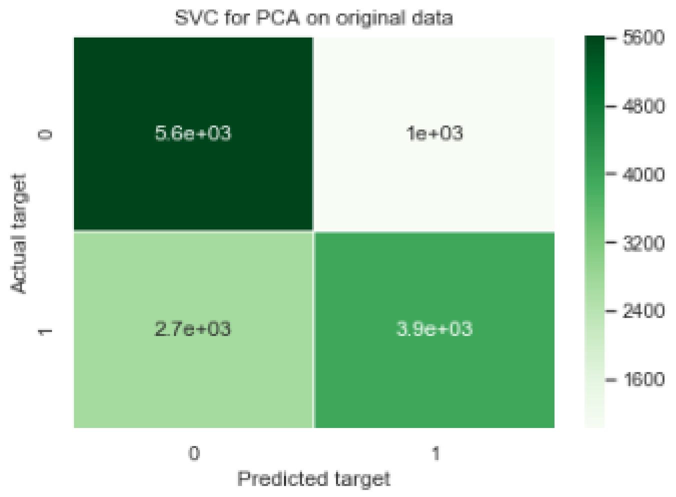


Predicting Default Payments Cont..

SVC Model

ii. With PCA on original data.

- ✓ Average accuracy = 70.4%
- ✓ Root mean squared error = 52.8%

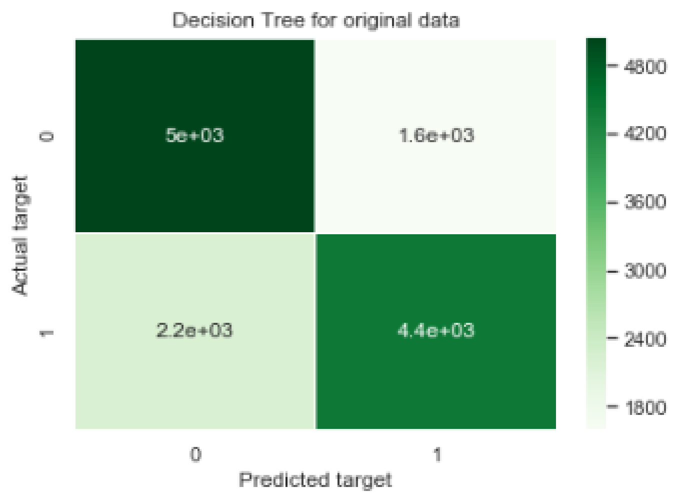


Predicting Default Payments Cont..

Decision Tree Model

i. With original data.

- ✓ Average accuracy = 69.3%
- ✓ Root mean squared error = 53.4%

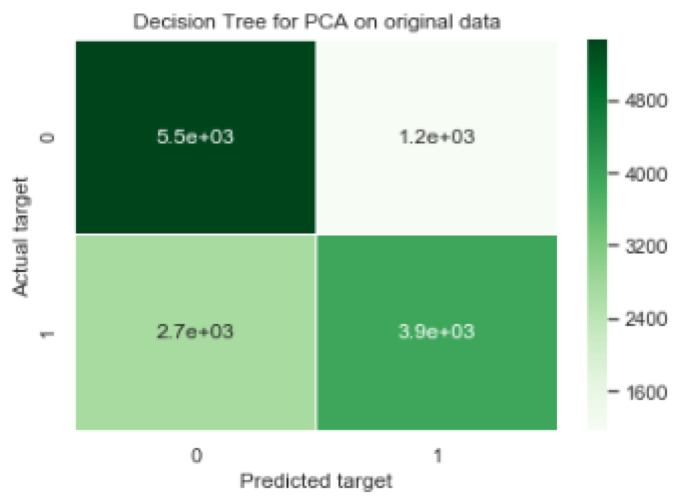


Predicting Default Payments Cont..

Decision Tree Model

ii. With PCA on original data.

- ✓ Average accuracy = 67.8%
- ✓ Root mean squared error = 53.9%



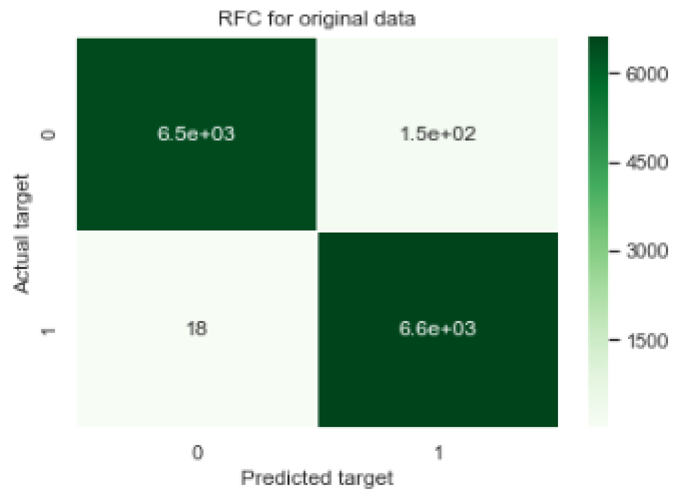
Predicting Default Payments Cont..

Random Forest Model

i. With original data.

✓ Average accuracy = 70%

✓ Root mean squared error = 11.1%



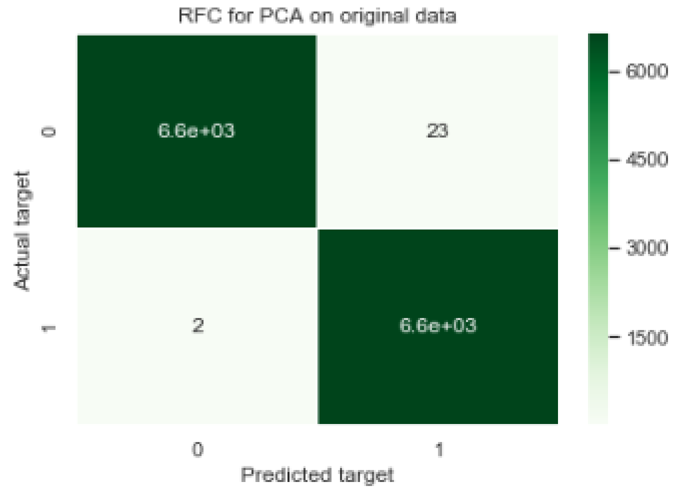
Predicting Default Payments Cont..

Random Forest Model

ii. With PCA on original data.

✓ Average accuracy = 70%

✓ Root mean squared error = 4.3%

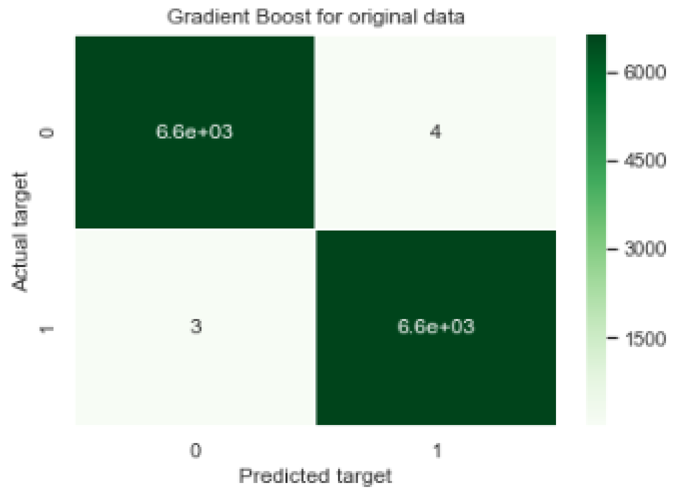


Predicting Default Payments Cont..

Gradient Boost Model

i. With original data.

- ✓ Average accuracy = 68.3%
- ✓ Root mean squared error = 2.2%



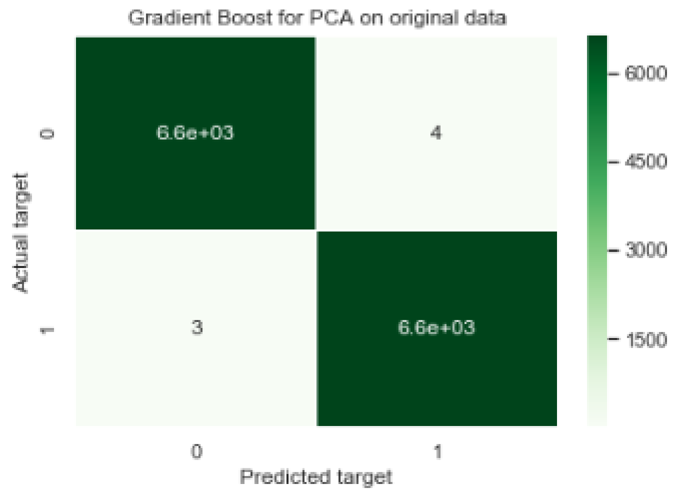
Predicting Default Payments Cont..

Gradient Boost Model

ii. With PCA on original data.

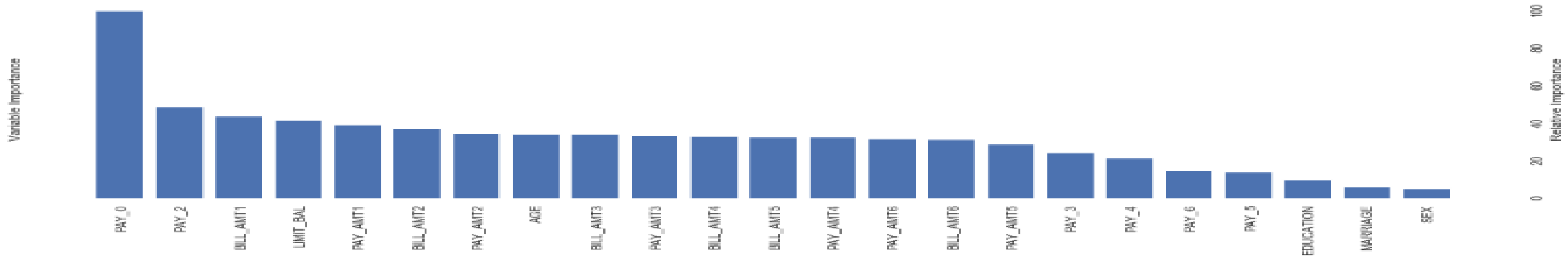
✓ Average accuracy = 68.5%

✓ Root mean squared error = 2.2%

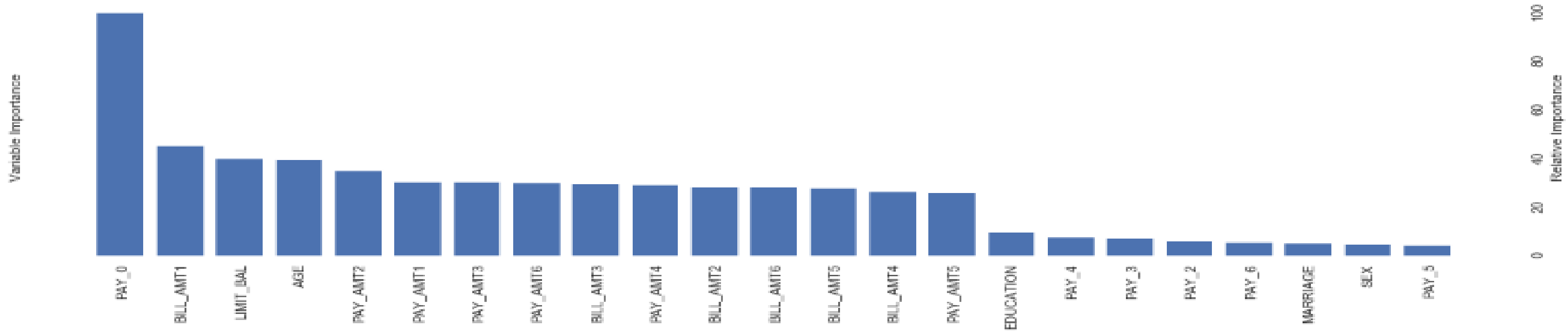


Feature Engineering

RFC important features



GB important features



Feature Engineering Cont..

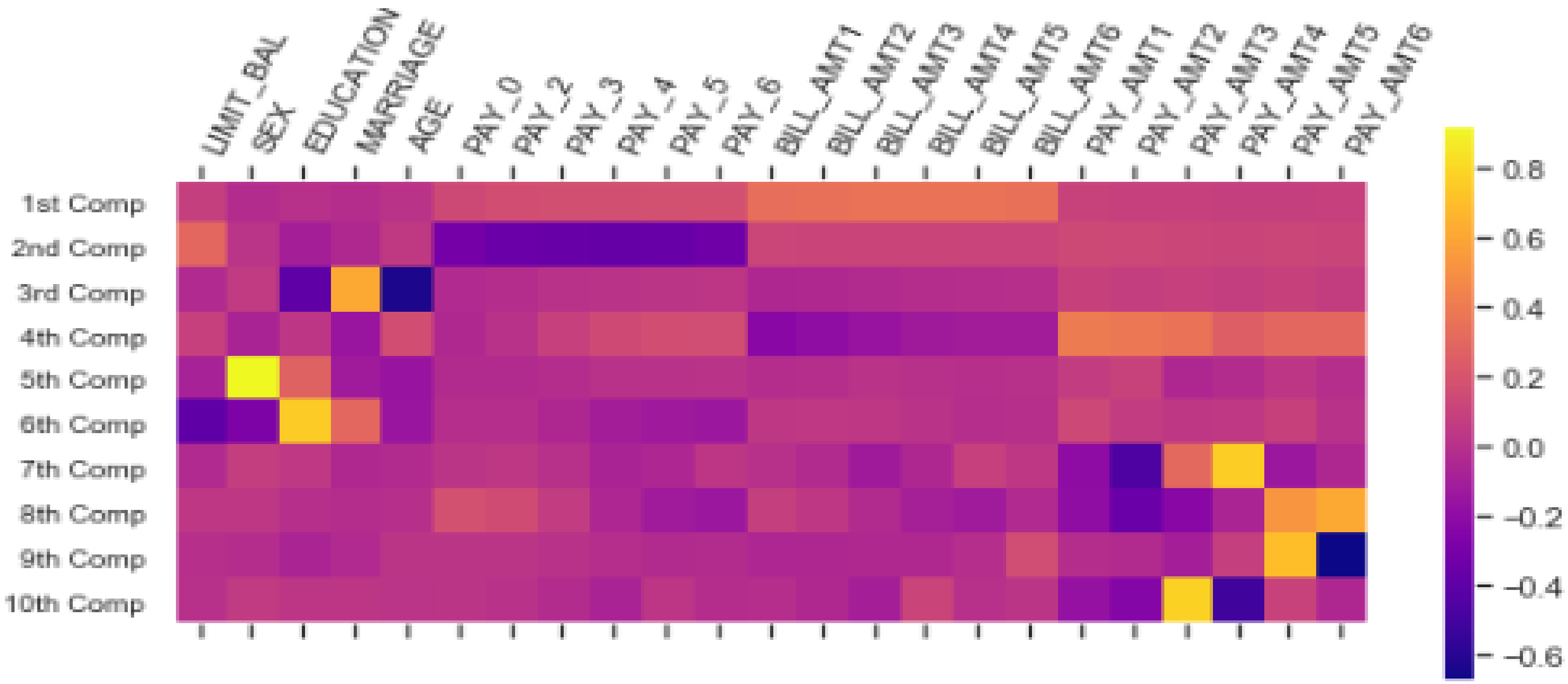
- Bill Average =

$$\frac{BILL\_AMT1+BILL\_AMT2+BILL\_AMT3+BILL\_AMT4+BILL\_AMT5+BILL\_AMT6}{6}$$

- Payment Average =

$$\frac{PAY\_AMT1+PAY\_AMT2+PAY\_AMT3+PAY\_AMT4+PAY\_AMT5+PAY\_AMT6}{6}$$

Feature Engineering Cont..

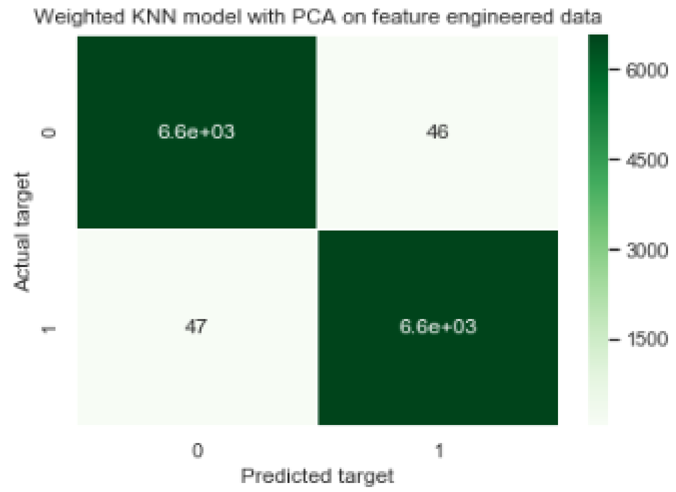


PCA Comp vs Feature Comp

Iteration and Evaluation

KNN Model

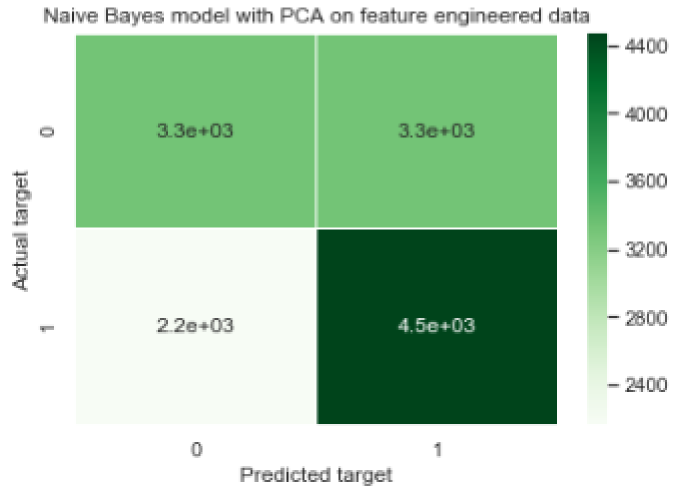
- ✓ With important features and feature engineering for Bill amount and payment amount
- ✓ Average accuracy = -3.3%
- ✓ Root mean squared error = 6.6%



Iteration and Evaluation Cont..

Naïve Bayes Model

- ✓ With important features and feature engineering for Bill amount and payment amount
- ✓ Average accuracy = -3.3%
- ✓ Root mean squared error = 64.2%



Best Prediction Model

	Score
SVC_pca	0.704793
RandomForest_data	0.695753
RandomForest_pca	0.695224
DecisionTree_data	0.693267
GB_pca	0.685053
GB_data	0.683621
NaiveBayes_data_train	0.678875
DecisionTree_pca	0.678647
SVC_data	0.515220
KNN_Weighted_data_train	0.082446
KNN_Unweighted_pca	0.017610
KNN_Weighted_pay_ind	0.007039
KNN_Weighted_Features	-0.033608
Naive_Bayes_Features	-0.033608
KNN_Weighted_pay_amt	-0.045792
KNN_Weighted_demo	-0.078221

	RMSE
GB_RMSE_data	0.022966
GB_RMSE_pca	0.022966
RandomForest_RMSE_pca	0.043401
KNNw_RMSE_features	0.066048
SVC_RMSE_data	0.071579
RandomForest_RMSE_data	0.111500
KNNw_RMSE_pay_amt	0.146593
KNNw_RMSE_demo	0.349620
KNNuw_RMSE_pca	0.388025
KNNw_RMSE_pay_ind	0.471973
SVC_RMSE_pca	0.528284
DecisionTree_RMSE_data	0.534945
DecisionTree_RMSE_pca	0.539573
KNNw_RMSE_data	0.598030
Naive_Bayes_RMSE_features	0.642631
NaiveBayes_RMSE_data	0.713472

Conclusion

- I have used about 6 different classifier models for predictions of this dataset. For every model used, I have tried different techniques in order to try to make our accuracy score better while trying to decrease our error rates.
- From the above list of Root Mean Squared error and accuracy score we can say 4 out of 16 techniques yield best values.
- I think I will choose Random Forest with PCA on original data as my best model. This is because of the lower root mean squared value compared to that of original data.
- I see that for this particular dataset, the ensemble models like Random Forest and Gradient Boost gave the best accuracy and after tuning the parameters (quite a few times), I could reduce the error rates drastically.

Next Steps..

- For this particular dataset, since the data was imbalanced and since I had 30000 rows, I have used random under sampling to balance the data.
- In the future I would try to sample more to balance data. Also, I think I would come up with more features and try to analyze the effect of the features on the different models I have used and try to continuously iterate and evaluate the models.

Thank you!