Customer Default Identification

Sugitha Devarajan

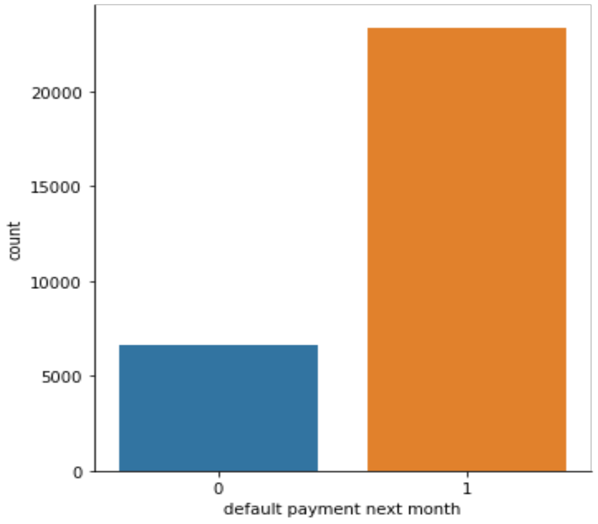
Problem

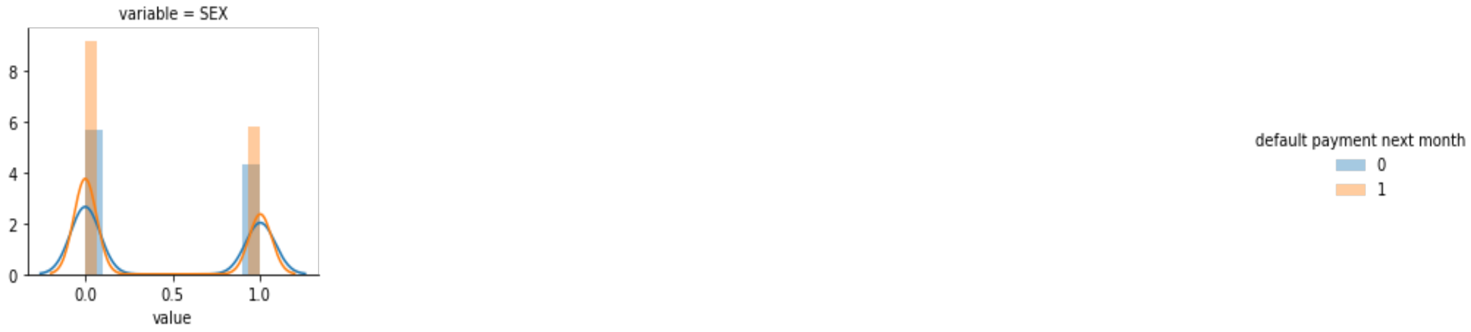
- An increase in customer default rates is bad for Credit One since its business is approving customers for loans in the first place. This is likely to result in the loss of Credit One's business customers. Need to build a model that can better predict what credit limit a customer should be assigned.
- The bottom line is Credit One need a much better way to understand how much credit to allow someone to use or, at the very least, if someone should be approved or not.

Questions to Resolve

- How do you ensure that customers can/will pay their loans?
- Can we approve customers with high certainty?
- How much credit to allow someone to use?

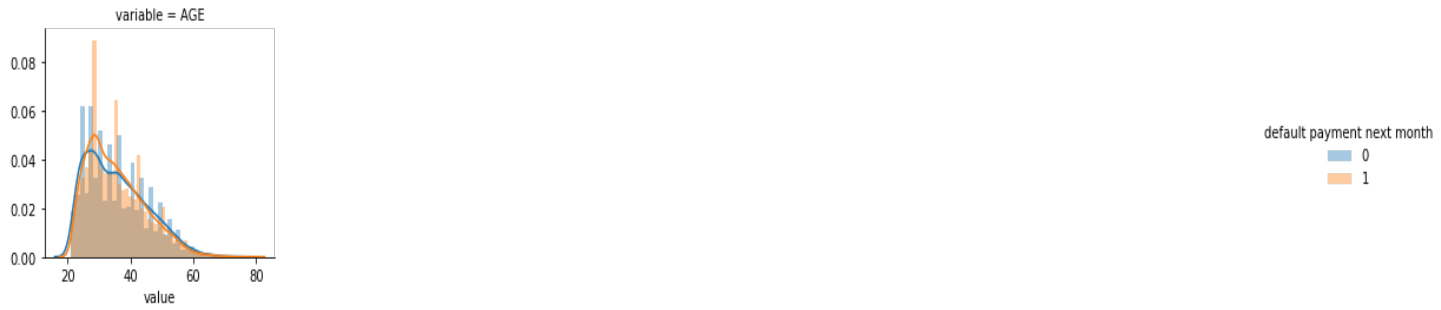
Few Observations during EDA





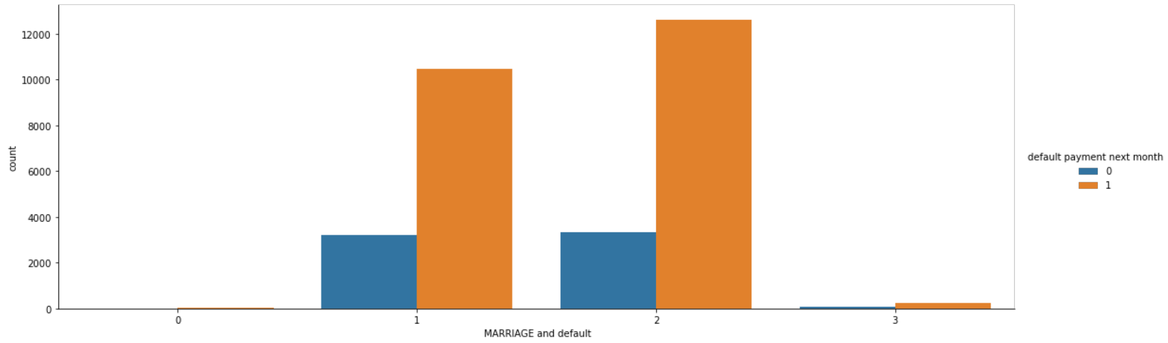
- We have an unbalanced data of default variable
- Will female demographics be more likely to pay the debt or male?

Not defaulted are more by females than males



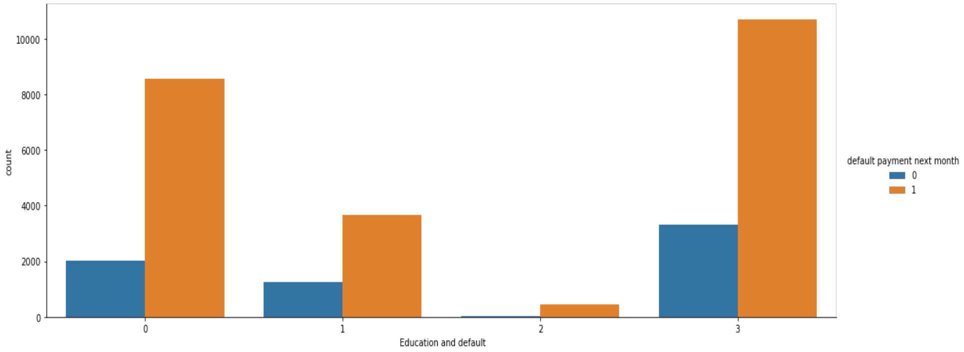
- What age group will likely to be defaulted?

Age 30 to 40 are high in not defaulting



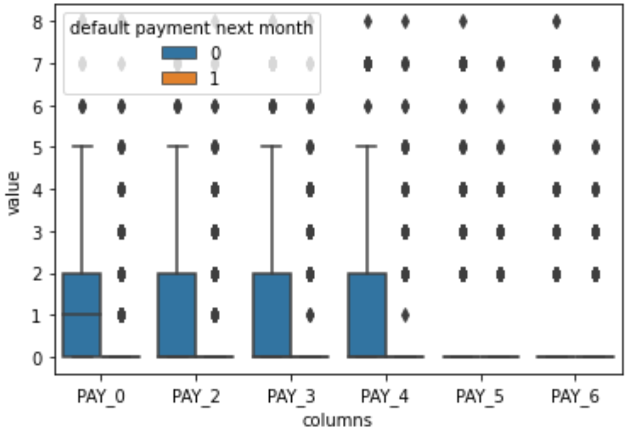
- Will marital status play a role in a customer becoming defaulted?

Single folks tend to not default more.



- Did education feature contribute more to the customer predicting to be defaulted?

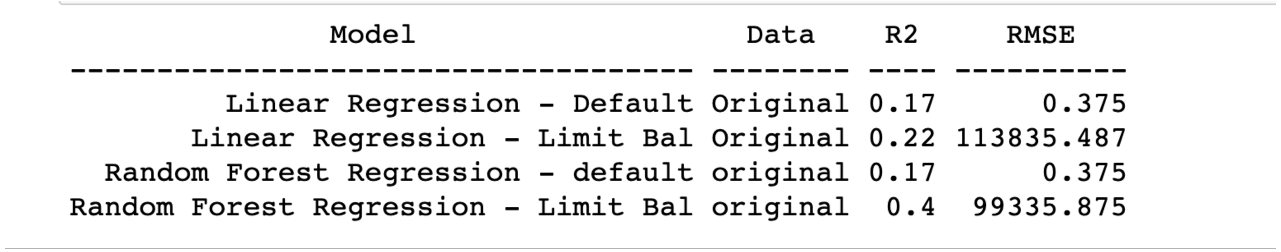
When customers are well educated then they don't default.



- Will history of payment let us know if a customer will be defaulted?

There is a tendency to default when they miss payments more than a month

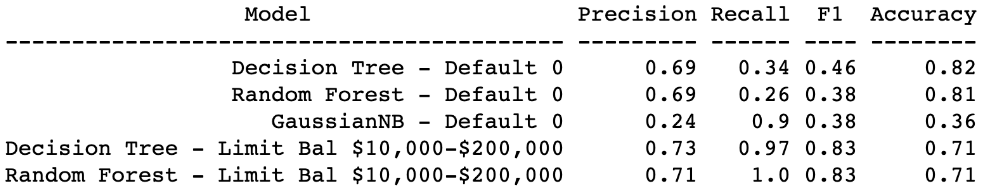
Credit Default - Machine Learning - Regression



```
              Model                     Data    R2     RMSE
------------------------------------ -------- ---- ----------
        Linear Regression - Default  Original 0.17      0.375
      Linear Regression - Limit Bal  Original 0.22 113835.487
 Random Forest Regression - default  original 0.17      0.375
Random Forest Regression - Limit Bal original  0.4  99335.875
```

- Since our aim is to predict the credit limit to be approved and to make sure customer does not default, I took the regression approach making the default and LIMIT_BAL as dependent feature to model and find the R2 and RMSE score.
- Lower R2 values correspond to models with more error, which in turn produces predictions that are less precise. In other words, if your R2 is too low, your predictions will be too imprecise to be useful. Conclusion - with regression model we were not able to predict default or answer the important question i.e., how much credit to allow. Moving on to classification

Machine Learning – Classification



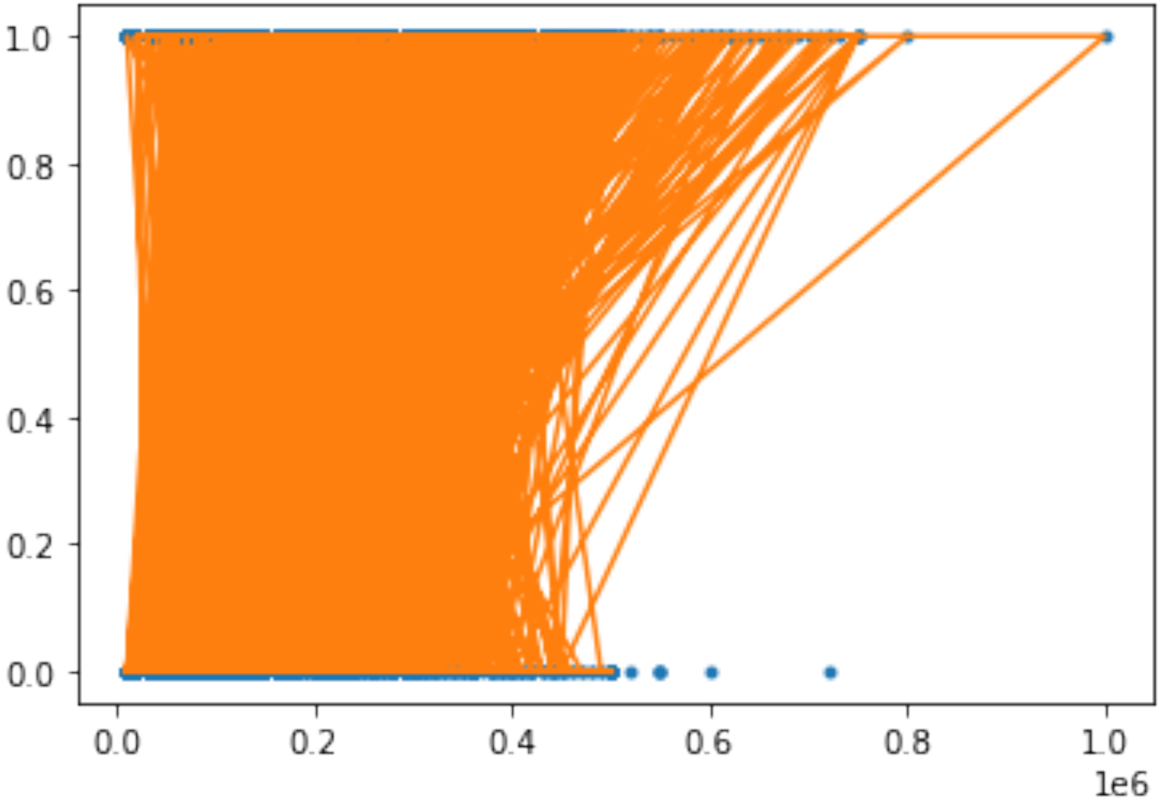
```
                  Model                    Precision Recall  F1  Accuracy
------------------------------------------ --------- ------ ---- --------
                 Decision Tree - Default 0      0.69   0.34 0.46     0.82
                 Random Forest - Default 0      0.69   0.26 0.38     0.81
                    GaussianNB - Default 0      0.24    0.9 0.38     0.36
Decision Tree - Limit Bal $10,000-$200,000      0.73   0.97 0.83     0.71
Random Forest - Limit Bal $10,000-$200,000      0.71    1.0 0.83     0.71
```

Accuracy: How often the model predicts the default and non default

Precision: When the model predicts default. How often it is true?

Recall: The proportion of the actual defaulters that the model will accurately predict.

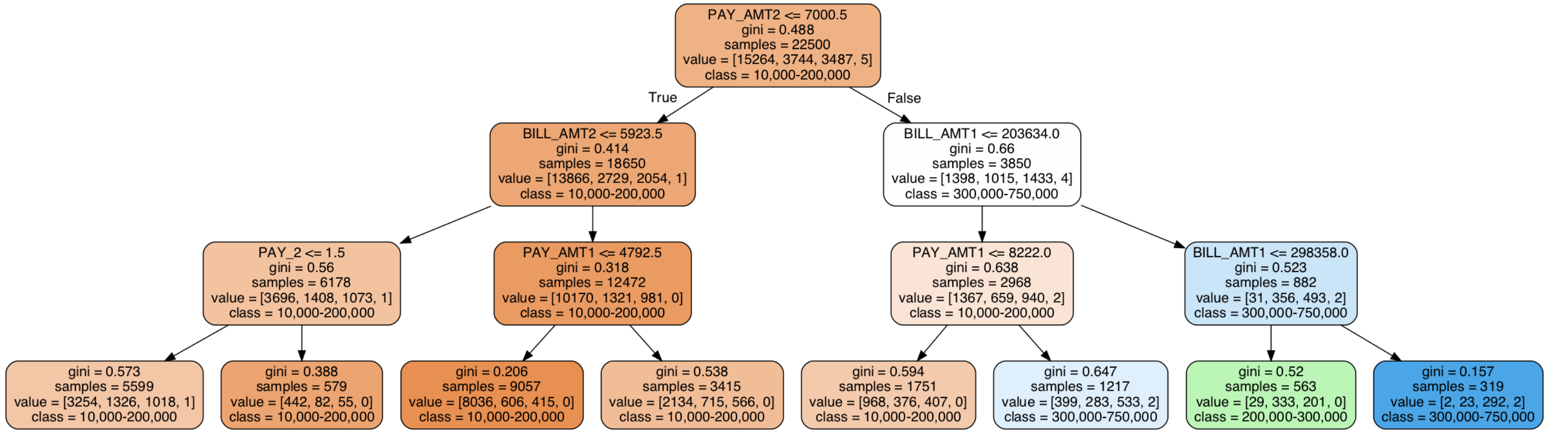
To answer the two main questions, I found the recall metrics in classification_report will help predict the ratio of True positives by True positives + False negatives. A false positive will hurt the cost of credit-one so, the GaussianNB seems to give 90% recallscore in predicting the default. And the decision tree algorithm helps predict the credit limit up to 97% recall when approving for between $10,000 and $200K.



- The model predicts the limit_bal to default more when its between 10,000 and 500K. Meaning higher limit balance are less likely to default.

Default

Credit balance



PAY_AMT2 <= 7000.5
gini = 0.488
samples = 22500
value = [15264, 3744, 3487, 5]
class = 10,000-200,000

True

False

BILL_AMT2 <= 5923.5
gini = 0.414
samples = 18650
value = [13866, 2729, 2054, 1]
class = 10,000-200,000

BILL_AMT1 <= 203634.0
gini = 0.66
samples = 3850
value = [1398, 1015, 1433, 4]
class = 300,000-750,000

PAY_2 <= 1.5
gini = 0.56
samples = 6178
value = [3696, 1408, 1073, 1]
class = 10,000-200,000

PAY_AMT1 <= 4792.5
gini = 0.318
samples = 12472
value = [10170, 1321, 981, 0]
class = 10,000-200,000

PAY_AMT1 <= 8222.0
gini = 0.638
samples = 2968
value = [1367, 659, 940, 2]
class = 10,000-200,000

BILL_AMT1 <= 298358.0
gini = 0.523
samples = 882
value = [31, 356, 493, 2]
class = 300,000-750,000

gini = 0.573
samples = 5599
value = [3254, 1326, 1018, 1]
class = 10,000-200,000

gini = 0.388
samples = 579
value = [442, 82, 55, 0]
class = 10,000-200,000

gini = 0.206
samples = 9057
value = [8036, 606, 415, 0]
class = 10,000-200,000

gini = 0.538
samples = 3415
value = [2134, 715, 566, 0]
class = 10,000-200,000

gini = 0.594
samples = 1751
value = [968, 376, 407, 0]
class = 10,000-200,000

gini = 0.647
samples = 1217
value = [399, 283, 533, 2]
class = 300,000-750,000

gini = 0.52
samples = 563
value = [29, 333, 201, 0]
class = 200,000-300,000

gini = 0.157
samples = 319
value = [2, 23, 292, 2]
class = 300,000-750,000

The lower the amount of given credit limit of the balance owing, the bigger the chances to default.

The DTC model predicts with recall score of 97% the lower limit making sure the default chance is lower.

Conclusion

- How do you ensure that customers can/will pay their loans?

When their credit limit is high the chances for a customer to pay their loan is high. When lower credit limits the chances are less.

- Can we approve customers with high certainty?

There are ways where our model can approve customers with certainty i.e., probability of default using recall score. In this model we focus on the default dependent variable and predict with 90% score recall whether a customer will default this way the cost of predicting false positive is to a minimum.

On other hand, there are safer ways with less risk demographics like female between age 30 -40 and education graduate or university who is single will have high chance of not being default.

Also, having a delay, even for 1 month in any of the previous months, increases the chance of default which the DTC model clearly proves.