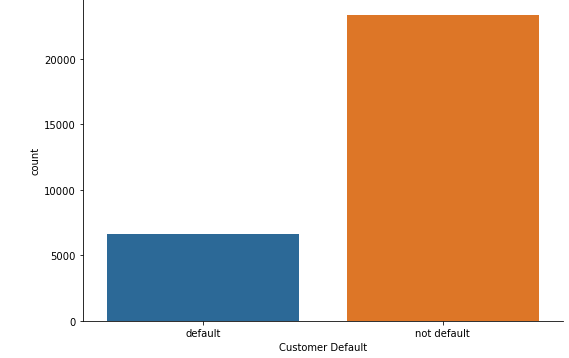
**Customer Default Identification Report**

**Problem**: Credit One seeing an increase in the numbers of customers who have defaulted on their loans. This is likely to result in the loss of Credit One's business customers and hence loss of profits.

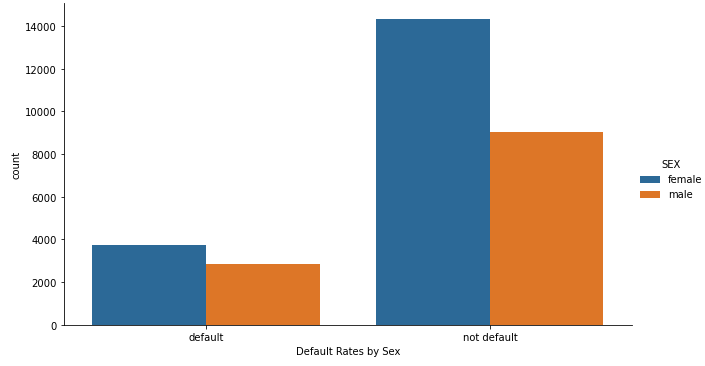
As shown below, of the data given, 22.12% of Credit One’s customers default on their payments.



**Data**

The majority of customers captured in this data are individuals with a university level education, females, and married and single customers.

**Default rates assessed by gender:**



Although the dataset has shown to be female heavy, we do see that females are more likely to default them men. However, as shown by the chart above, this does not seem to be a significant data point.

**Default rates assessed by pay period:**

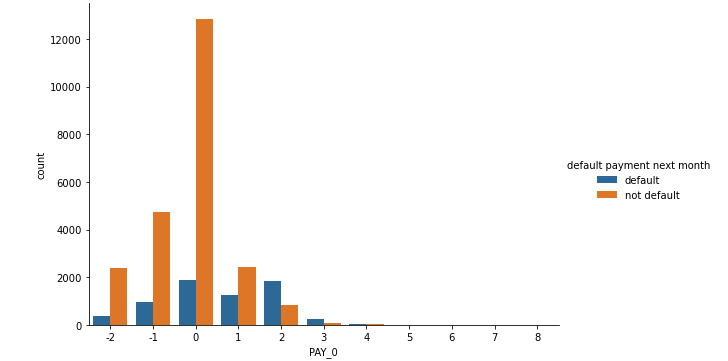
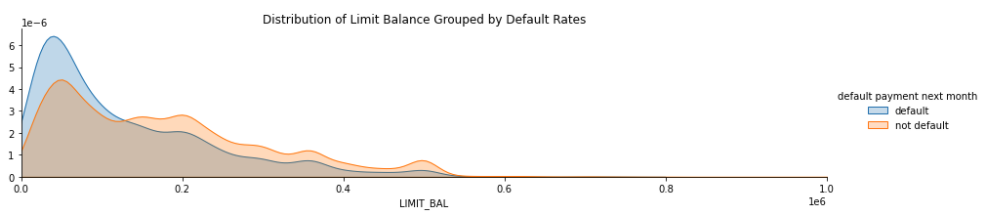


Chart above shows that individuals are most likely to start defaulting on payments if they delay by 2 months.

**Distribution of limit balance grouped by default rates:**



Data shows highest default rates when lower limit balance. Overall, shows trend of defaulting decreasing as limit balance increases.

Marital status, education and age did not have a significant impact to note on default rates.

**Questions to investigate:**

1. How do you ensure that customers can/will pay their loans?
2. Can we approve customers with high certainty?

We were tasked with building a regression model that can better predict what credit limit a customer should be assigned. Regression models used: Random Forest Regression, Linear Regression and Support Vector Regression. Here we used ‘LIMIT\_BAL’ as our dependent variable and based on the EDA, we altered the independent variables to select the pertinent features. As mentioned above, our EDA suggested that marital status, age and various education levels and did not have a significant impact on default rates and so these were not originally included as our independent variable. However, after experimentation, I was able to conclude that we achieved higher accuracy and prediction scores with those variables included. Having said this, our accuracy scores still remained low with the Random Forest Regressor and Linear Regression algorithms scoring ~0.1%. In addition to experimenting with the independent variables, I also tested the models based on a limited dataset i.e., on data where customers did not default and still achieved low scores.

Although we were aiming to build a regression model that can better predict what credit limit a customer should be assigned, we were not able to achieve success using ‘LIMIT\_BAL’ as our dependent variables in a regression model. I began thinking if this problem would be better be answered if viewed as a classification problem. When ‘LIMIT\_BAL’ was cut into 5 bins were able to achieve ~0.68% across all classification models (Random Forest Classifier, Decision Tree Classifier and Gradient Boosting Classifier).

Upon further analysis, I thought instead of creating a model to predict a customer’s limit balance, that it may be more useful and effective to the business to create a model, which when assigning limit balance, can predict (based on customer demographic data) if a client is likely to default on their payments. Therefore, we can input a limit balance and run it through the model and see that a customer is likely to default on their payments if we assign *xyz* as their limit balance. I explored this by assessing classification models using 'default payment next month\_default' as the dependent variable as was able to achieve an average of ~0.78 across the classification models.

To conclude, we cannot always ensure that a customer can/will pay their loans due to unforeseen circumstances e.g., an epidemic; however, but we can build a predictive model to assess that in normal times, whether a customer is likely to default on their loans, and which individual attributes are likely to result in an individual defaulting on their loans.

**Recommendations**

**Data limitations:** In the future, datapoints to consider capturing should include examples of the below:

* Gross annual income
* Monthly payment obligations e.g., rent
* Debt-to-income ratio
* How individuals limit balance changed over period of time e.g., do we see a pattern of limit balance increasing as customer make their payments and then customers becoming unable to make payments