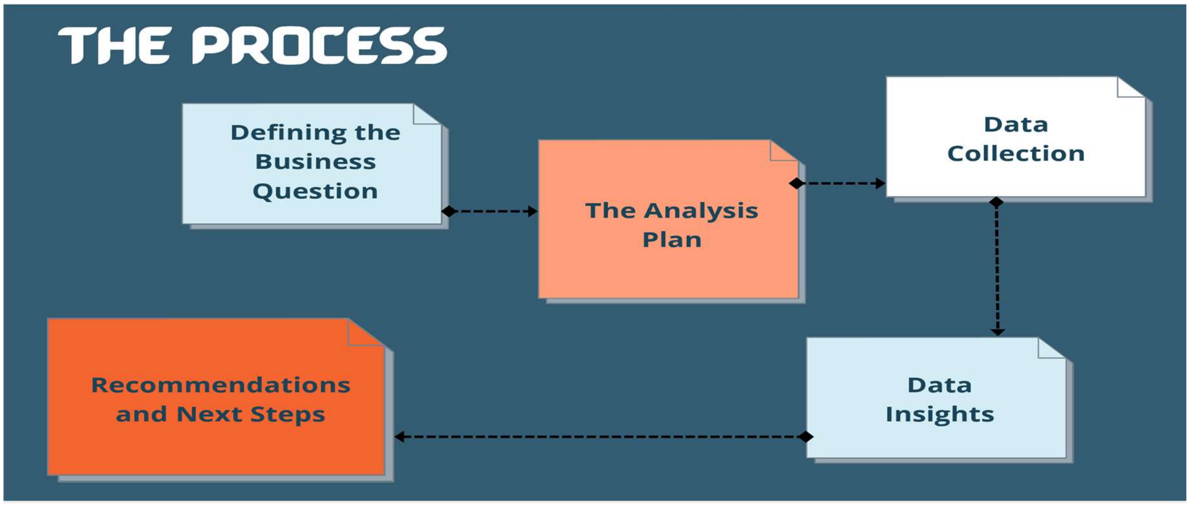
Credit One Project

Final Report

The goal of this project was to utilize data science to address the critical business issue of increased customer defaults which threatens the business of Credit One.

The following was the process used by the data science team to address the task:


THE PROCESS
Defining the Business Question
The Analysis Plan
Data Collection
Data Insights
Recommendations and Next Steps

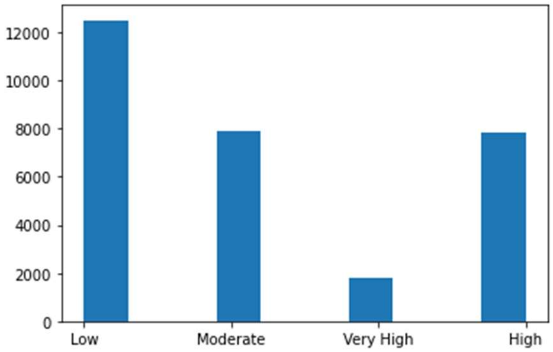
- ✓ Recap of the **Business Question**:
  - ➲ Credit One needs a much better way to understand how much credit to allow someone to use or, at the very least, if someone should be approved or not.
- ✓ Recap of the **Analysis Plan**:
  - ➲ Goal: To examine current customer demographics to better understand what traits might relate to whether or not a customer is likely to default on their current credit obligations.
  - ➲ Hypothesis: Which customer attributes relate significantly to customer default rates and can a predictive model be used to better classify potential customers as being 'at-risk'?
  - ➲ Required Data: Customer Demographics, Payment History, Credit Limit, Default Status
  - ➲ Methodology: Data descriptives and correlations. Machine learning regression and classification methods in Python.
- ✓ This report will address **Data Insights** and **Recommendations/Next Steps**.

Data Insights

- **EDA**: An extensive exploratory data analysis (attached) was conducted and revealed several observations that were communicated in a previous report. To summarize:
    1. The overall current default rate is about 22%
    2. Bill Amounts and Payment Amounts have very similar distributions within, indicating that customers tend to have consistent behaviors.
    3. Overall, default status did not seem to have a strong relationship with any demographic or variable other than payment history.
    4. There is a positive correlation between Credit Limit and Education-graduate school. This means that customers with higher degrees are more likely to get higher credit limits.
    5. A negative correlation between credit limit and payment history indicates that the higher the credit limit, the more likely customers pay the balances. However, this is affected by outliers.
    6. Some of the outliers were removed. Discretization alleviated this issue when running classification machine learning models.

- **Regression Models**: Since all variables were converted to numerical, regression models were tested and evaluated.
    1. First ***Default Status*** was used as the predicted variable.
    2. Multiple types of regression algorithms were assessed, including *Random Forest Regressor, Linear Regression,* and *Support Vector Regression.* However, none yielded reliable evaluation metrics.
    3. Feature engineering was experimented with to see if it will produce better metrics, however, the regression model remained unreliable.
    4. Next, ***Limit Balance*** was used as the predicted variable.
    5. Again, multiple types of regression algorithms were assessed recursively, and although it was much better than the Default Status dependent variable, still none yielded enough reliability.
    6. Feature engineering was experimented with here as well t, however, the regression model remained insufficiently reliable.

- **Classification Models**: To assess classification algorithms, ***Limit Balance*** was discritized into four categories consisting of: Low, Moderate, High, and Very High.
  7. Using ***Limit Balance Categories*** as the predicted variable, classification algorithms were assessed recursively, including *Random Forest Classifier, Decision Tree Classifier,* and *Gradient Boosting Classifier*.
  8. The best algorithm was Gradient Boosting Classifier, however, with an accuracy of 62%, its reliability remains insufficient.



Recommendations

- **Machine Learning Models**: Can they reliably predict whether a customer will default or not? Our recommendation is that using the current demographics and payment history, machine learning models ***are not*** a reliable predictor.
- There could be other sources of data that will be much stronger predictors such as customer income, occupation, and net assets.
- However, the current demographics are weak predictors.
- Additional recommendations: The insight that customers with higher degrees receiving higher limit balances needs to be further examined. Is it because people with higher degrees are more likely to have better qualifications for higher credit? Or is it an unsupported anecdotal assumption that people with higher degrees are more likely and/or able to pay back funds?

- **Initially Posed Questions to Investigate**:
  1. *How do you ensure that customers can/will pay their loans?*
     Given the current data, it is not possible to ensure that.
  2. *Can we approve customers with high certainty?*
     No, there are no statistically significant predictors to provide such certainty.