





Data Science Framework Report

Credit One
Tasneem Dawoodjee



Credit One

- **Who We Are:** Credit One, a third-party credit rating authority
- **What We Do:** Provide retail customer credit approval services to businesses
- **Our Problem:**
 - Increase in customer default rates
 - **If this continues, we will lose business**
- **Current Solution:** Our processes are no longer correctly identifying customers who will default
- **Objective:** Use data science to build a model to correctly identify customers who will default on loans

Data Science Framework

1. Define the Goal
2. Collect and Manage Data
3. Build the Model
4. Evaluate and Critique the Model
5. Present Results and Document
6. Deploy and Maintain the Model

Description and Location of Related Data

- Data consists of csv file with 24 attributes from 30,000 customers from April 2005 to September 2005
 - Amount of Credit, Gender, Education, Marital Status, Age, Past Payment Histories, Bill Statement Amounts, Previous Payment Amounts, Default Status
- Issues with out data:
 - Some attributes have large range of values which we will address by using binning techniques
 - These include age, bill statement amounts, payment amounts, and loan amounts which for example vary from 10,000 to 1,000,000.

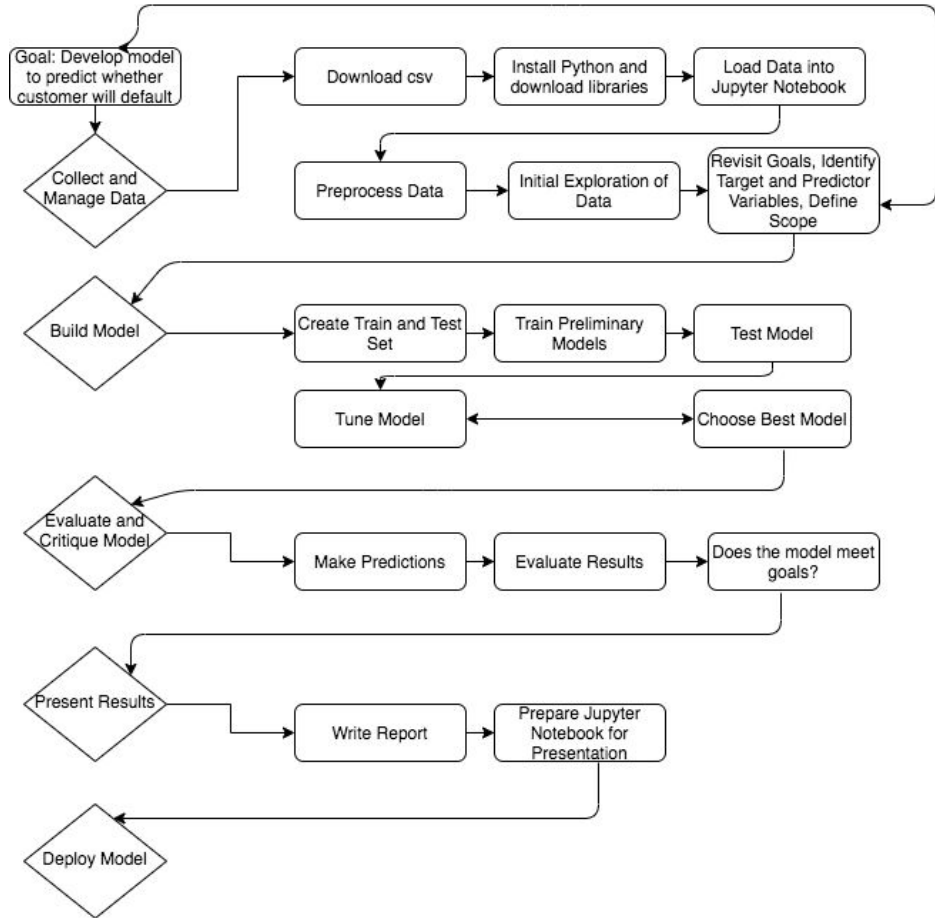
Issues with Data

- Education: 7 levels (0 to 6) but only 3 defined: 1 for graduate school, 2 for university, 3 for high school. We will condense 0, 4,5,6 into one category “other”
- Lack of Longitudinal Data.
 - Payment Histories: We have no record of missing payments longer than 8 months.
- Missing Data: Date of Default
 - There are customers with zero (or less) bill amounts and zero (or less) payments that still defaulted. Helpful to know when customer defaulted on loan. It is not possible for a customer to have no payments and even credits in their account, yet still default on a loan. There may be attributes we have not included in the dataset, or bill and payment histories that are outside of the date range collected in this dataset.
 - This anomaly represents 1% of the dataset. We will remove customer with these kinds of irregularities from the dataset, especially if they represent less than 5% of the customers.

How We Will Manage the Data

- Data Source: original unaltered csv file is with Guido Rossum, Senior Data Scientist, Credit One
- Data Security:
 - Data has been de-identified. No customer names, or location details. Customer names have been replaced with an identification number from 1 to 30,000.
- Programming Language: Python
- Platform: Jupyter Notebook
- Libraries: NumPy, Pandas, Matplotlib Python 2D, Sci-Kit Learn

Our Data Science Process



Initial Insights and Recommendations

1. Class imbalances can affect predictive capability of our models.
 - a. 23,364 customers in our dataset, or 78% did not default. 6,636 customers, or 22%, did default.
 - b. 2,873 out 11,888 males defaulted. 3,673 out 18,112 females defaulted. We have significantly more females in our dataset than males.
 - c. If we have more data with additional customers that defaulted or males, we should include it while developing our model.
2. Loan start dates and loan end dates can be helpful attributes to include in dataset. Because we don't have these attributes, we don't know how much relevant payment histories are. The data does not tell us whether some loan agreements allow customers to pay off their loans at different time intervals.