Data Science Framework Report

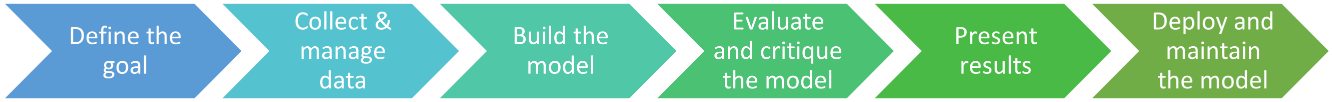
Chapter 5, Task 1
By: Radhika Ghosh

The goal

The goal is to use data science and Python and assist Credit One in identifying which customers may be at a risk of defaulting on their loan payments.

Proposed Data Science Framework

I propose the following Data Science Framework defined by Zumel and Mount. This framework provides a 6 step approach, the first of which is to understanding what the stated goal is, i.e. identifying customers that may default on loans. Second step, is to collect and manage the data which in this is case is the Credit One customer data. The third step is to build a set of data science predictive models that would analyze the customer attributes that have been provided in order to predict weather a certain customer would default on his/ her loan. The fourth step compares the various predictive models and identifies the one likely to perform best on unknown data. In the fifth step, we would present these findings to Credit One and in the final step we would deploy the predictive model and set up maintenance to alter and update the model as required in the future.



Data Sources

The data for customer default history will be provided by Credit One. Our team requests a data scrubbed of client identifying information. This dataset would be used solely for the data science process and be stored in a secure data warehouse. Our proposed method using Python can connect to these data sources securely.

Data Management

The data will be stored in a secure data warehouse and the deployed data science model will establish a secure connection to the data warehouse.

Known Issues with Data

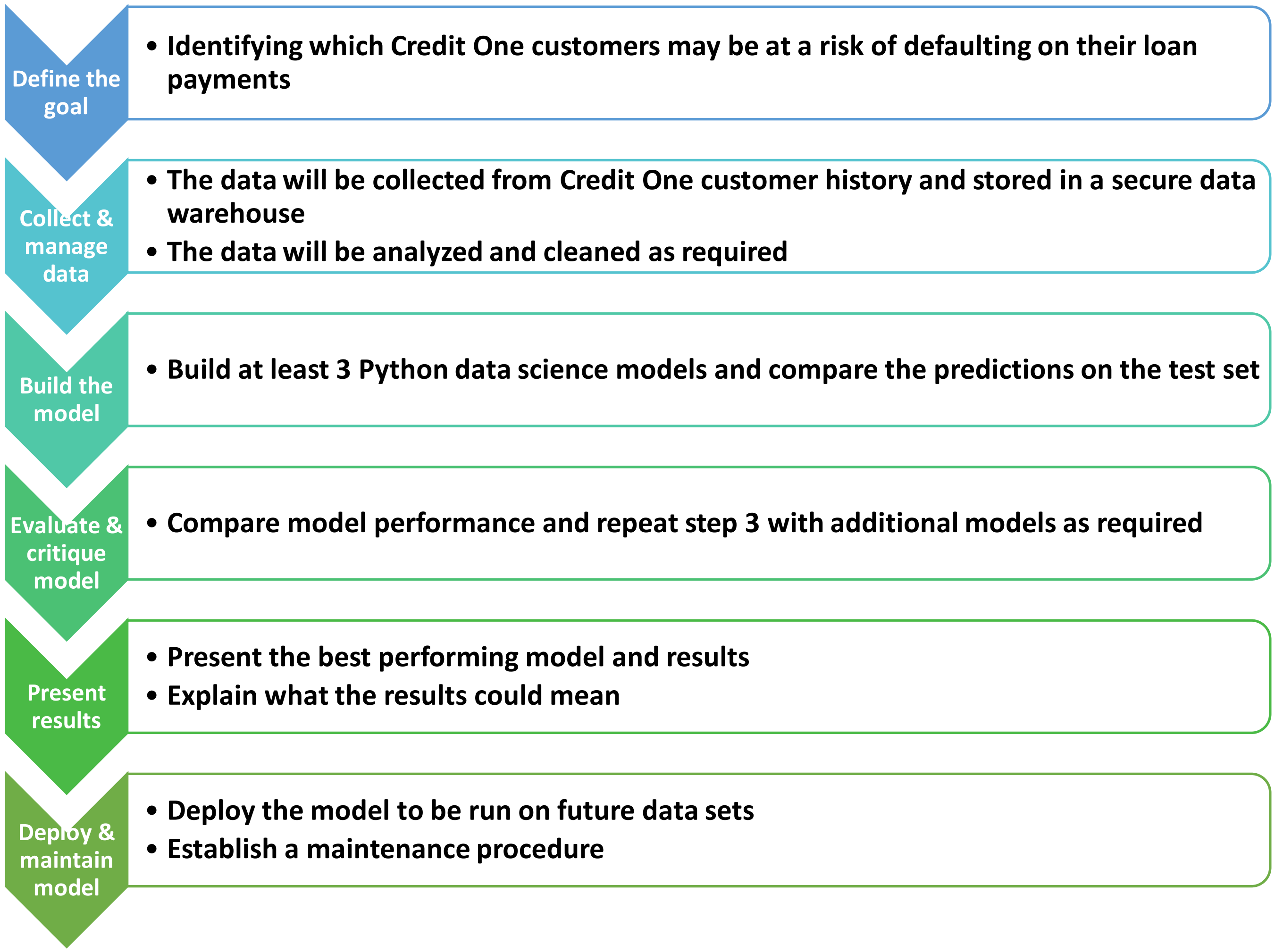
The following issues were noted with the data set that has been provided:

1. The column headings are inconsistent and do not provide a lot of insight. These will be updates during the data science process

2. There are some inconsistent records, such as those that show the client has had “no consumption” of their credit card and thus has made no payments, is listed as having defaulted. It is not clear from the data if they had a previous remaining balance before they stopped using their credit cards. These records would require some analysis to ensure they are not invalid data.

Data Science Process

The following data science process based on the Zumel and Mount data science framework will be used for this project.



Define the goal

- **Identifying which Credit One customers may be at a risk of defaulting on their loan payments**

Collect & manage data

- **The data will be collected from Credit One customer history and stored in a secure data warehouse**
- **The data will be analyzed and cleaned as required**

Build the model

- **Build at least 3 Python data science models and compare the predictions on the test set**

Evaluate & critique model

- **Compare model performance and repeat step 3 with additional models as required**

Present results

- **Present the best performing model and results**
- **Explain what the results could mean**

Deploy & maintain model

- **Deploy the model to be run on future data sets**
- **Establish a maintenance procedure**

Initial Insights

An initial review of the data showed the following:

1. 22% of the customers in the sample data appear to have defaulted
2. Most customers have the payment status of 0 or the use of revolving credit
3. There doesn’t appear to be any missing data

4. The data for Payment status is split across multiple columns and will likely require to be discretized.