C5T1

Informal Report

John Miller

Goal

- To develop a creative, empirically sound solution for reducing the number of defaults among customers of partners Credit One serves. This should be achieved using a data science method, and mitigate losses of client revenue

Data Science Process

- Preprocess data - assess issues in data types, needs to center/scale, discretization, etc
- Exploratory Data Analysis - use visualization and summary functions to begin to shape the data and learn the data set and identify potentially productive questions and hypotheses
- Develop Hypothesis - based on EDA results, develop a formal hypothesis to be the basis for developing a model
- Develop Model - develop various models and tune parameters to maximize desire results
- Test Model - test developed model to assess for accuracy, error, or other measures of confidence
- Assess Results and Confidence - assess results of first round of models, and determine if additional modeling, new data sources, etc. are needed to improve results. Identify if the hypothesis is inadequate, and revise if needed.

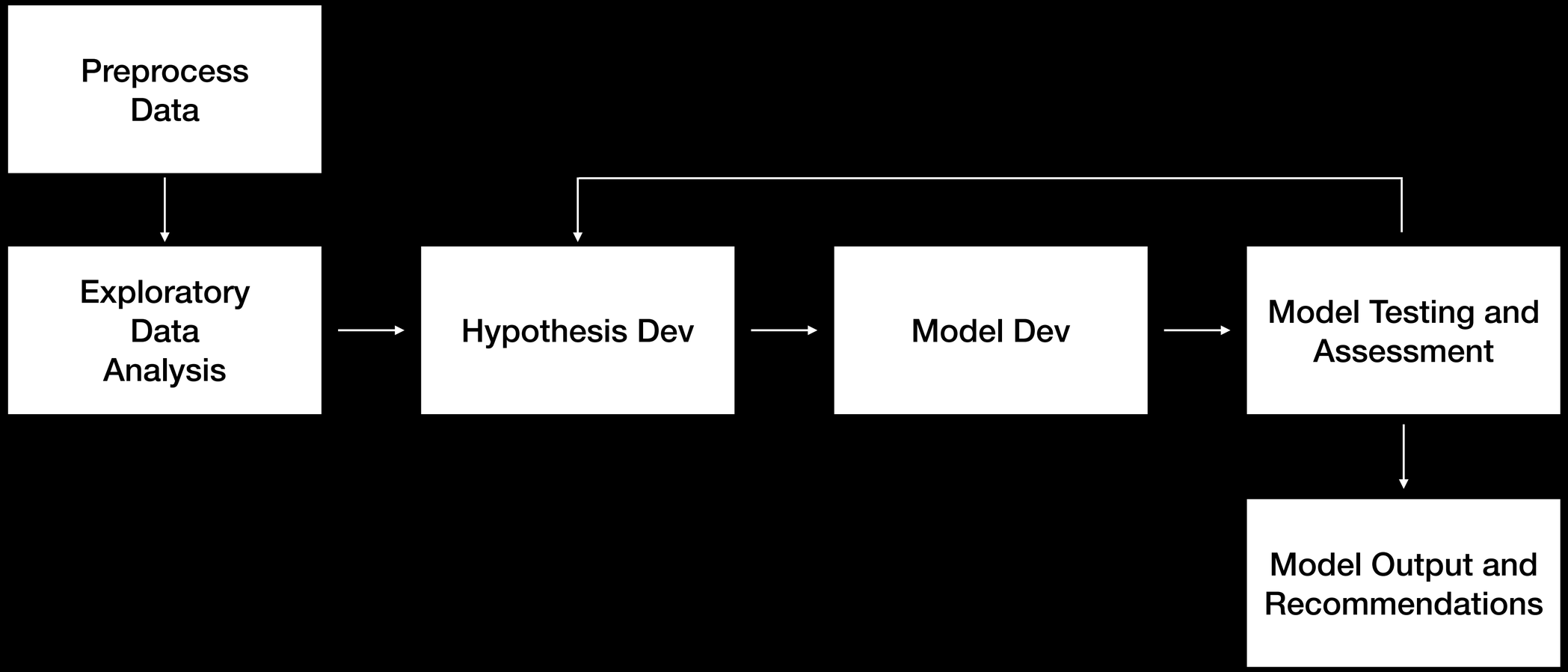
Data Location + Management

- Data Provided by Guido
- Will be managed with local storage and GitHub

Known Issues

- Some data will need to be recoded (-2 and -1 for repayment statuses, e.g.)
- Some data will need to be discretized (age)
- Some data may be wise to scale

Process Flowchat



Initial Insights

- What role do demographic features such as education and marital status play in default likelihood?
- Will it be useful to look at unpaid balance as an integer of dollars, or as a percentage of total credit
- Can a data set like this have household income data added?