Credit One

Analysis of Customer Defaults

Proposing R Framework for Analysis

Practical Data Science with R (Zumel and Mount)

- Define the goals
- Collect and manage data
- Build the model
- Evaluate and critique the model
- Present results and document
- Deploy and maintain the model

Why R Framework?

- Method starts with clear understanding of business goals
- Efficient data analysis remains focused on business goals
- Actionable insights achieved through sound data science structure

Definition of goals

- Why do stakeholders want to do the project?
  - Credit One has seen an increase in customers defaulting on loans in the past year
    - Currently 22% of total customers have defaulted on loans
  - Rising defaults leads to revenue loss for business clients and potential loss of clients for Credit One
- **Statement of goals** (what do Stakeholders' need)?
  - Identify and understand differences in customer demographics by default versus no default
  - Find out if we can predict credit limits with 80% or higher level of certainty of not defaulting
- Why is current solution inadequate?
  - Defaults with status quo are increasing and Credit One risks losing clients the longer problem is left unsolved

Definition of goals

- What resources do you need?
  - Project team: Stakeholder, Senior Data Scientist, Data Science lead, Operations manager
  - Credit One historical customer data
  - Python data analysis software and associated libraries
  - Kick-off and mid-point project meeting, weekly touch bases with Senior Data Scientist
- How will the results of your project be deployed?
  - If goals are achieved, predictive model will be deployed by Data Science and Operations team
  - Enhancements and bug fixes to occur weekly following model deployment

Collect and Manage Data

- What data is available?
  - Historical dataset of credit card customers
- Initial data insights
  - 30,000 total customers
  - Includes prior 6 months of customer info
  - 25 total variables include customer demographics, monthly billing & payment information, and default status
  - Customer demographics include sex, education, marital status, age
  - 78% of customers have not defaulted, 22% of customers have defaulted on loans
- Will it help to solve the problem? Is it enough?
  - Further analysis will be needed to determine if the data will be sufficient to solve the problem

Collect and Manage Data

- Is the data quality good enough?
  - Initial review shows no missing or duplicate data
  - Further analysis is needed to determine if data quality is good enough to achieve goals
- Are there any known issues with data? If so, how will they be addressed?
  - Variable names can be changed to make more understandable
  - Must convert variables with word values to number values so software can analyze data
  - Remove unnecessary rows (header definitions) and columns (ID) irrelevant for data analysis
- **Exploratory data analysis (EDA)** will be conducted to:
  - Identify and understand differences in customer demographics by default versus no default
  - Identify relationships between variables to determine datapoints most useful for modeling

Build the model

- What techniques might I apply to build the models?
  - A minimum of 3 models will be built to determine best performing model
  - Feature selection (keeping impactful variables) will occur to optimize model performance
  - Models will be tuned to enhance accuracy
  - Models will undergo validation steps to minimize error
  - Models will be evaluated to determine if accurate enough to meet stakeholder needs

Evaluate and critique the model

- Is the model accurate enough to meet stakeholders' needs?
  - Stakeholders' want to reverse the trend of rising customer defaults
  - Current data reveals that 22% of Credit One customers have defaulted on loans
  - Model accuracy of 80-90% is generally considered successful
  - Further analysis is needed to determine if model meets stakeholders' needs
- Does it perform better than the obvious guess?
  - Further analysis is needed to determine if models perform better than obvious guess
- Do the results of the model make sense in the context of the real-world problem domain?
  - To be determined

Present Results and Document

Present key findings and recommendations to stakeholders by May 23, 2020

- How should stakeholders interpret the model?
- How confident should they be in its predictions?
- When should they potentially overrule the model's predictions?

Simple Outline

- Objective
- Background
- Scope
- Approach
- Recommendations
- Key insights with impact
- Next steps

Deploy and Maintain the Model

Implement process to ensure model runs smoothly

- How is the model to be handed off to "production?"

- How often, and under which circumstances, should the model be revised?

Visualization of R Framework (Zumel & Mount)

Potential pitfalls (and solutions)

- Business goals are unclear (ask questions in the beginning to clarify)

- Data quality not good enough (recommend additional data variable collection)

- Data is not good enough to meet goals (revisit project design and goal defining stage)

- Model does not solve problem (return to data collection/management and model building stage)

- Recommendations are vague (recommendations should be actionable and have positive impact)