Data Science Framework Report

Jennifer Zhu
October 20, 2019

Agenda

Background of the Project

Objective/Goals of the Project

Data Science Process Framework

Data Management

Data Description

Any Known Issue

Flowchart Visualizating the detailed process

Insights

Background

Credit One has seen an increase in the number of customers who have defaulted on loans they have secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away. This project is to help Credit One design and implement a creative, empirically sound solution.

Statement of the Goal

A data science project using python to determine problems behind the loss of business caused by customers defaulting on loans and to find sound solutions.

Data Science Process Framework

- Define a measurable and quantifiable goal
- Collect and manage Data
- Build the model
- Evaluate and critique the model
- Present the result and document
- Deploy and maintain the model

Management of Data

Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users.

We have elected to use Python and a few different libraries to do the heavy lifting for us.

- Set up an environment with Jupyter Notebook format
- Hosted on GitHub
- Use of numpy, pandas, matplotlib, Scipy, Sci-Kit Learn

Description of Data

The data that we are going to use comes from a research aimed at the case of customers default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods.

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others).

X4: Marital status (1 = married; 2 = single; 3 = divorce; 0=others).

X5: Age (year).

X6 - X11: History of past payment. Past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005.

Description of Data

The measurement scale for the repayment status is: -2: No consumption; -1: Paid in full; 0: The use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Y: client's behavior; Y=0 then not default, Y=1 then default"

Any Known Issue with the Data

The data we have now can not provide us with a good picture of the overall capacity of the clients - the ability to pay back his/her loan

What we know about the Client

- Amount of the given credit
- Amount of bill statement
- History of past payment
- Amount of previous payment

What we don't know about the Client

- How much credit given compared to the gross monthly income
- How much debt compared to gross monthly income
- How much payments compared to monthly gross income

Flowchart

Data Science Process Flowchart					
Define a Goal	**Collect & Manage Data**	**Build the Model**	**Evaluate & Critique the Model**	**Present the Result & Document**	**Deploy & Maintain the Model**
• Why do the stakeholders want to do the project? • What do they need from it? • Why is their current solution inadequate? • What resources do you need? • How will the result of your project be deployed?	• What data is available? • Will it help to solve the problem? Is it enough? • Is the data quality good enough?	• Which techniques might I apply to build the model? • How many techniques should I apply?	• Is the model accurate enough to meet the stakeholders ' needs? • Does it perform better than "the obvious guess" and any techniques being used currently? • Do the results of the model make sense in the context of the real-world problem domain?	• How should stakeholders interpret the model? • How confident should they be in its predictions? • When should they potentially overrule the model's predictions?	• How is the model to be handed off to "production" ? • How often, and under which circumstanc es, should the model be revised?

Flowchart - Reasons

A straightforward goal is set - Stop losing business caused by defaulted loan

A previous research data is available - predictive accuracy of probability of default

Research method and tools are elected - Python and its various libraries

Insights

Although we don’t know a client’s monthly gross income, but the following could also help to understand a client’s capacity in order to find a sound solution to project:

Amount of credit compared to amount of payment

Amount of credit compared to amount of bill statement

Amount of bill statement compared to amount of payment

Thank you!