ITM 6899

Capstone Project

Student Name(s):

Advisor: Dr. Chongqi Wu

Date:

Overview & Purpose

Units: 1

Development and writing of business analytics project.

The purpose of the project is to practice the knowledge obtained during the study of the business analytics program with a specification in classification problem in Machine learning. The project will use a fabled company and dataset to demonstrate a real world business problem, after the course, student will learn in a real industry classification problem, the classification project cycle and related techniques, algorithms, metrics, evaluation methods and so on.

Remarks and Requirements:

1. **Must enroll in ITM 6899 Capstone Project with Dr. Chongqi Wu to receive grade for the Capstone Project.**
2. **Familiar with Python and Jupyter notebook.**
3. **Students may work in group of up to 4 students.**
4. **Submit a project report which addresses all the eight questions at the end of this file. Project report must also include one page of executive summary.**
5. **Submit your Jupyter notebook along with the report.**
6. **Use this file as the cover of your project report.**

Objectives

1. To get familiar with data exploration steps in a Machine Learning problem
2. To practice the visualization methods of different feature analysis
3. To learn the typical type of machine learning problem - binary classification and its model building cycle, as well as evaluation metrics
4. To learn the importance of deal with the missing data, outlier in the dataset
5. To get a understanding of what is an imbalanced data classification problem

Business Problem Description

Beta is an online e-commerce company. The company is interested to know in an early stage, after their customer convert to a paid customer, whether they could become a VIP consumer of their website or not within a month (30 days). The have a dataset where is observed and aggregated during their first 7 days since the first date they made their first purchase. The dataset and its features are explained as below. Once they have the classifier, they could target those VIP customers with personalized treatment.

The Task

Use this dataset to build a binary classifier to use the first 7 days of data since a customer convert, whether they will become a VIP customer for the business within 30 days since their first conversion. The definition VIP by day 30 conversion is defined as a customer spend equal or more than $500 by day 30.

Dataset Description

1. IsVIP_500 : target variable, class label, 1 means is a VIP by day 30, 0 means not.
2. payment_7_day : total payment made by day 7 of conversion
3. dau_days: distinct days of customer login to the website.
4. days_between_install_first_pay: number of days since the user registered on the website
5. total_txns_7_day: total transactions the customer made on the website in the first 7 days.
6. total_page_views: number of product items the customer viewed on the website in the first 7 days.
7. total_product_liked: number of product items they have marked like during their views in the first 7 days
8. product_like_rate: the products liked divided by viewed products
9. total_free_coupon_got: number of free coupons the customer got during the first 7 days after conversion.
10. total_bonus_xp_points: total bonus points customer got during the first 7 days, where they could use it as cash with certain redeem rate.

Task Break Down and Submission Requirements

Please use jupyter notebook and python as program language to answer the following questions. One may use text, code, plot to support your answers.

1. Describe the data with statistical information about each feature.
2. Visualization of each feature and the target variable (class distribution).
3. Is there any missing data in the dataset? If so, how to deal with it? What are the methods to deal with missing data?
4. Is there any outlier in the data set? If so, how to deal with it? What are the methods to deal with outliers?
5. Show how to build the classifier and how you will evaluate the results? What is the metrics you will use?
6. What algorithms you plan to use and why? (at least 2 or more)
7. How would compare the results of the algorithms? Which is better, why?
8. Given the current model, how would you improve the results? The data is imbalanced, please give a few method that the academic will use for imbalanced data, pick one to experiment with your current dataset and see how the result will change.