UDACITY MACHINE LEARNING ENGINEER NANODEGREE CAPSTONE PROJECT

@alwzhang | April 3, 2022

OVERVIEW

This project report is a detailed and comprehensive showcase of the Capstone Project for Udacity's Machine Learning Engineer Nanodegree. The goal of the report is to explain the process and stages of the project development and share the findings.

1. Define the Problem

> *Define the problem you want to solve and investigate potential solutions and performance metrics.*

Starbucks was established in Seattle, Washington in 1971, by three students who attended the University of San Francisco. When the first Starbucks was open, it was a small shop in the Seattle's historic Pike Place Market, selling specialty coffee beans and freshly brewed coffee. Since then, Starbucks has quickly expanded into a global enterprise that is well known for its coffeehouses, where customers can enjoy its authentic beverages, food, packaged coffee, and more. As of 2019, Starbucks was ranked 121st in the list of Fortune 500.

One of Starbucks key marketing strategies is sending promotions and advertisements via Starbucks mobile app. Customers can order coffee for pickup, load money and pay from the app, and collect rewards points. Promotional offer and bonus points are incentives that drive more orders/sales.

The goal of sending customers promotions and offers is to increase their purchases. However, it is not a desirable strategy to send all offers to all customers at the same time. The goal of the project is to take advantage of the transactions and demographics data to determine the offers that should be targeted to different groups of customers.

This project goal can be achieved by binary classification machine learning models, such as Logistic Regression, Random Forest, Neural Networks etc. Performance metrics include model accuracy, precision, recall, and F-1 score.

2. Analysis

> *Analyze the problem through visualizations and data exploration to have a better understanding of what algorithms and features are appropriate for solving it.*

Data Exploration, Data Cleaning, and Feature Engineering

- Closer look at original datasets
  - Portfolio
    - Encode channels and offer_type features
  - Profile
    - Rename id to customer_id
    - Engineer became_member_on (date) feature to days_as_member (int)

        - Encode gender feature
        - Keep notna rows
    - Transcript
        - Rename person to customer_id
        - Engineer value feature, break it down to offer_id, amount, reward
        - Create a successful_offer feature as a label/target feature
            - Multiply offer viewed and offer completed to determine if the offer was successful and triggered the transaction (1 = successful offer)
            - This is an important feature engineering step as some offers were completed but have not been viewed
- Merge and format all above to get final data and save as master_df.csv
    - Encode offier_id to offer_1 to offer_10
    - Drop customer_id
    - Make successful_offer the first column for X, y data split
    - Drop email feature after evaluating correlation matrix
- Final features
    - Target: successful_offer
    - Features: difficulty, duration, reward, mobile, social, web, offer_type_bogo, offer_type_discount, offer_type_informational, age, income, days_as_member, gender_F, gender_M, gender_O, offer_1, offer_10, offer_2, offer_3, offer_4, offer_5, offer_6, offer_7, offer_8, offer_9

Statistics and Visualizations

- Details can be found in the notebook

3. Implementation

Implement your algorithms and metrics of choice, documenting the preprocessing, refinement, and postprocessing steps along the way.

There are some popular binary classification algorithms that will be useful to this problem. In this project, the following algorithms had been implemented:

- Logistic Regression
- Random Forest
- Gradient Boosting Machine
- Support Vector Machine
- Neural Networks

Preprocessing steps include:

- Split independent feature and dependent features
- Train test split with 80/20 split
- Scale the data using different scaling methods

Postprocessing steps include:

- Calculate model's accuracy score
- Calculate model's F-1 score
- Plot confusion matrix

4. Results

Collect results about the performance of the models used, visualize significant quantities, and validate/justify these values.

Statistics and visualizations can be found in the notebook.

Model	Accuracy	F-1 Score
Neural Networks	0.775952	0.717305
Random Forest	0.780850	0.703379
Gradient Boosting Machine	0.774293	0.702489
Logistic Regression	0.757703	0.686305
Support Vector Machine	0.754780	0.682553

5. Conclusion

Construct conclusions about your results and discuss whether your implementation adequately solves the problem.

Since this is a binary classification model, F-1 score will be a better choice as it is the weighted harmonic mean of recall and precision. By comparing model performance, Neural Networks and Random Forest are two final winners.

An ensemble model can be potentially helpful, transforming to a weighted average approach for the final predictions. Additional model tuning, for example, by using grid search, might improve the model performance. Introducing more features will also be an interesting route to explore and compare or combine with the current set of models.