Machine Learning Engineer Nanodegree

Udacity

Capstone Project

Proposal

Shanmukha Priya Mudigonda

Table of Contents

Project: Predicting article retweets and likes based on the title using Machine Learning

You can read the full project on this file and check here the implementation

This is the final project of the specialization Machine Learning Engineer Nanodegree.

Abstract - This project derives from the direct marketing system which Starbucks uses to keep in touch with its customers. Aiming to incentivize and reward the customers registered in its platform, Starbucks periodically sends individual messages containing offers related to its products. There are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational. In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount. In a discount, a user gains a reward equal to a fraction of the amount spent. In an informational offer, there is no reward, but neither is there a requisite amount that the user is expected to spend. Offers can be delivered via multiple channels: e-mail, social media,

on the web, or via the Starbucks's app. Nonetheless, marketing campaigns have associated costs. Hence, to be considered a successful campaign, it must generate profit higher than that initial cost. That means, companies expect to have a return on investment (ROI) as high as possible. Thus, companies do not want to spend money sending offers to customers that are not likely to buy their products. On the other hand, new customers need to be attracted, so it is necessary to identify who are the people with a higher probability to respond to a marketing campaign. Sometimes, recurrent consumers deserve some reward so they can feel appreciated and not forgotten. However, some customers keep coming back even if they do not receive offers. To give an example, from a business perspective, if a customer is going to make a 20 dollar purchase without an offer anyway, you would not want to send a "buy 20 dollars, get 4 dollars off" offer, unless this relative short-term loss means a more satisfied customer who will consume more in the future. Another case is those customers who only buy products when receiving some reward, while other ones are opposed to marketing campaigns and do not want to be contacted at all. Those are a few examples that illustrate how complex is the marketing decision process that has been faced by companies for years. Considering the recent advances of artificial intelligence and the massive amount of data gathered over the years, this is a topic that could be widely improved by intelligent systems because they can analyze a large amount of data and understand patterns sometimes hidden for the human perception. With data related to Starbucks, it was used machine learning methods including support vector machines (SVM), decision trees, gaussian model (Decision Tree, Gaussian NB and Random Forest) to make the predictions. This study shows that the MultinomialNB model performed better for retweets reaching an accuracy of 60.6% and logistic regression reached 55.3% for likes.

Keywords - prediction, machine learning, social media, title, performance

Dependecies

This project requires **Python 2.7** and the following Python dependencies installed:

- NumPy
- Pandas
- matplotlib
- Jupyter Notebook
- scikit-learn
- Anaconda
- Nltk

Run

In a terminal or command window, run one of the following commands:

```
ipython notebook title-success-prediction.ipynb
```

or

```
jupyter notebook title-success-prediction.ipynb
```

This will open the Jupyter Notebook software and project file in your browser.

Note

The Capstone is a two-staged project. The first is the proposal component, where you can receive valuable feedback about your project idea, design, and proposed solution. This must be completed prior to your implementation and submitting for the capstone project.

Motivation

The motivation behind this project was exploring the Starbuck's Dataset which simulates how people make purchasing decisions and how those decisions are influenced by promotional offers. It was driven by 4 main questions that were answered as the project came to its completion:

- What is the proportion of client who have completed the offers based on Gender?
- What is the proportion of client who have completed the offers based on their Age?
- What is the proportion of client who have completed the offers based on their Income Level?
- What are the most important features that help drive the offers in customers?

Datasets and Inputs

For this project, the data sets are provided by Starbucks and Udacity in the form of three JSON files. These contains simulated data that mimics customer behavior on the Starbucks rewards mobile app.

- portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json - demographic data for each customer
- transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer.
- reward (int) - reward given for completing an offer.
- duration (int) - time for offer to be open, in days.
- channels (list of strings)

profile.json

- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account.
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

transcript.json

- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

Files

The following files attached to this GitHub's repository include the following:

- **Starbucks_Capstone_notebook.ipynb**: This is the Jupyter Notebook in which I performed all my work.
- **proposal.pdf**: This document contains the initial project proposal I submitted prior to necessarily beginning this project.
- **data**: This contains the three JSON files provided by Starbucks / Udacity as noted above.

IV. Solution Statement: To face the problem stated above, this project proposes to apply machine learning techniques to study customers' behavior by analyzing the transcriptions of their relationship with Starbucks. More specifically, a neural network will be trained to predict how customers may react when receiving each one of the available offers: if they will complete the offer cycle or not. So, it will be possible to identify which one is more suitable for each customer. Since consumers' behavior is not a feature isolated in the time, the next actions are affected by past experiences. Regarding this time-dependency, this project proposes to build and train a Recurrent Neural Network (RNN) as the central piece of the solution, aiming to analyze the customer behavior through the time. The result of this project is a direct marketing system that, given a customer, is able to predict the likelihood of each offer be completed.

V. Benchmark Model: A more traditional model will be trained in the same dataset used by the intended Recurrent Neural Network, so that the results are comparable. In this case, it means training a Feedforward Neural Network (FNN). Basically, an FNN analyzes a static input and makes predictions not considering the customer history. Differently, RNN is able to make predictions based on past events, instead of analyzing the current moment as an isolated situation.

A naive network model might understand an offer as adequate because it produced a good result in the past and tends to repeat that action every time. However, it may not be able to detect whether the same offer becomes inadequate when sent a second time to the same customer, perhaps because the customer does not want to repeat the same purchase forever. In another case, the customer is conditioned to buy products, so no offer sending is necessary anymore. However, that naive network keeps suggesting the same offer over again. This relationship between past experiences and future behavior is what the Recurrent Neural Network is supposed to recognize. Building and training both models allows us to compare the predictions made considering only the static user state (FNN) and those made based on the customer history (RNN). Then, we will be able to evaluate whether the problem stated is better addressed with a Recurrent Neural Network model.

VI. Evaluation Metrics: The accuracy of the models will be measured to evaluate the performance of the networks. By using the same metric, we can quantify and compare both the benchmark and the final models. Considering that customers might have variations in their standard behavior, having an accuracy very close to 100% might indicate that the network just memorized the customers' behavior instead of understanding their consumption patterns. On the other hand, having too low accuracy also might indicate that the network was not able to learn general patterns. Hence, the target accuracy for the RNN in this project is about 80%.6

VII. Project Design: The theoretical workflow for approaching the solution stated includes several machines learning techniques, following the guideline sections below.

a. Data loading and exploration Load files and present some data visualization in order to understand the distribution and characteristics of the data, and possibly identify inconsistencies.

b. Data cleaning and pre-processing Having analyzed the data, handle data to fix possible issues found.

c. Feature engineering and data transformation Prepare the data to correspond to the problem stated and feed the neuralnetworks. The transcription records must be structured and labeled asappropriate offer or not.

d. Splitting the data into training, validation, and testing sets Prepare three datasets containing distinct registers within each one. The largest dataset is employed to train the networks, while the validation set, to evaluate the models during the training phase. The testing set contains data never seen before by the networks, so it willbe possible to consider this dataset as being new interactions between Starbucks and customers. By using this dataset, it will be possible to measure the final performance and compare the results of the trained models.

e. Defining and training a Feed-Forward Neural Network Training of the benchmark model.

f. Defining and training a Recurrent Neural Network Training of the proposed model.

g. Evaluating and comparing model performances Comparison between the accuracy of both network models to verify each one is more suitable to solve the problem stated.

h. Presenting predictions for offer sending Present the resulting predictions, along with discussions on how this system should be employed

Summary

The project was completed in the following phases:

- Dataset Exploration and Preprocessing: - In this process, the data provided was cleaned, processed and merged for further analysis.
- Data Analysis & Visualization: In this process, few questions were answered by analyzing the question and providing visualization which answers the questions 1-3:

In order to compute the proportion of male or female within the age range of 29 to 69 with income ranging from 30k to 100k who have completed the offer, we need their respective data and the total number of participants in the test.

- Creation of supervised Model and Performance Evaluation: In this step, various machine learning model (Decision Tree, Gaussian NB and Random Forest) were tested on the given dataset and their respective performances were evaluated.
- Important Feature Analysis: In this process, the best estimator among the three was used to predict the most important feature that impacted the promotional offer completion in customers.

We found out that Time, Duration, Reward, Difficulty and Discount are the top 5 feature drivers that estimate the offer completion with Time being the highest impact driver for features.