Capstone Proposal-Starbucks Project

Machine Learning Engineer Nanodegree
Benjamin Redmond

Domain Background

This project aims to show how machine learning can be used to provide efficient targeted advertising. The advertising world is continuously becoming more digital and that brings about a vast amount of data that can be harnessed to provide powerful information for marketers. Advertising strategies have been researched extensively to improve the effectiveness of advertisements across various target audiences [1].

ML techniques enhance the accuracy of targeting by predicting the most relevant adverts for users based on pre-existing user data [2]. It is possible to utilise this information to provide a better return on advertising spend. More and more marketers are turning to this method to better decide which adverts will have a positive impact on which demographics.

Used in this way ML aims to take past information about customers responses to advertisements and make predictions and build models about customer behaviour in the future. It examines the historical data and detects patterns autonomously. These patterns are often hidden from humans attempting to analyse data without leveraging ML. The models can then be used to continuously improve marketing by refining its ability to influence the desired audience.

Problem Statement

This project uses simulated Starbucks data to see how customers react to adverts that are sent to their mobile app. The objective is to identify trends within the customer data that can be used to optimise the efficiency of the adverts sent. If we target certain adverts at customers who we know are more likely to be receptive to these adverts, then it will be possible to get a better return on advertising spend. It will also provide a better customer experience if we can stop sending adverts to customers who we think have little to no interest in them. Neither the company nor the customer want redundant adverts sent.

The adverts in this project can either be a simple advertisement for a drink or an actual offer such as a discount of BOGO (buy one get one free). Every offer has a validity period before expiring. Customers receive varying adverts at varying intervals.

Some basic demographic data is also provided about the users. We can also see the time and value of the transactions they made. We will also be able to see if they viewed the offer which is of relevance because the customer could have made a purchase without seeing the offer. The goal will be to identify the most appropriate offer to send to a customer such that they view the offer and then complete a purchase before the offer expires.

Datasets and Inputs

The datasets are provided in JSON files by Udacity. It contains simulated data that mimics Starbucks data as briefly described in the section above. Three files are provided:

- portfolio.json – Offer ids and meta information about each offer
- profile.json – Demographic data for each customer
- transcript.json – Records for transactions and information on offers viewed, received or completed

portfolio.json

It contains 10 rows and 6 columns.

- id (string) - offer id
- offer_type (string) - type of offer i.e., BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) – number of days that an offer is open
- channels (list of strings) – medium used (email, mobile, social, web)

profile.json

It contains 17000 rows and 5 columns. Although 2175 rows have age with value set to 118 which represents a missing value. The gender and income fields are also missing values on the rows where age is 118. I suspect these will not be useful during the analysis and will need to be filtered out.

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

transcript.json

It contains 306534 rows and 4 columns. Less than 11% of the rows are classified as “offer completed”. This may affect how I prepare the data or how I measure the effectiveness of the model.

- event (str) - record description (i.e., transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test

- value - (dict of strings) - either an offer id or transaction amount depending on the record

Solution Statement

Neural networks will be the core of this project. They will be trained using the Starbucks data to analyse how customers respond to receiving an advert. A recurrent neural network could prove to be very appropriate at this task because RNNs also take past results into account. It is possible a customer will respond either more positively or more negatively to repeated offers. Perhaps a customer gets irritated by continuous adverts or perhaps once a customer completes the offer the first time and enjoys the product and no further offers are needed as they would purchase the product going forward regardless. Alternatively, a customer might have enjoyed the offer and be waiting expectantly for a further offer before they purchase a Starbucks product.

This RNN machine learning technique will be used to see which offer should be sent each customer depending on their demographic data and historical behaviour in order to maximise the possibility they will make a purchase.

Benchmark Model

A feedforward neural network will be used as a benchmark against the RNN. An FNN differs from a RNN in that it does not consider historical behaviour. A FNN model might believe a particular offer to be the optimum one in a particular scenario because it worked well in the past but as explained above this might not necessarily be always the case.

Using a FNN as a benchmark against a RNN it will be possible to determine whether historical responses to adverts significantly influence future responses. We will be able to see if a RNN and its added complexity is the better neural network to use in this problem area on this dataset.

Evaluation Metrics

I will use accuracy to evaluate the effectiveness of the neural network models. Accuracy is the ratio of correct predictions to number of predictions. I believe this to be a more useful metric than precision, recall or F1-score in this situation. Precision and recall are important metrics when false positives and false negatives respectively are of higher importance (such as predicting fraud or cancer screening). In this task the consequence of having false negatives and false positives is not critical. I believe a high accuracy is what is most valuable here.

Although accuracy may not be the best metric if we have an uneven class distribution. At first glance this may be the case as less than 11% of the transaction values are classified as “offer completed”. It may prove more insightful to focus more on false positives or false negatives. The AUC-ROC method is highly optimal for a binary classification problem with imbalanced dataset classes as it plots true positive cases against false positive cases.

I will compare the accuracy of the RNN model against the accuracy of the FNN model to see if the RNN model performs better. I believe an accuracy percentage of 70% is a good score for an effective model. I will also measure the AUC-ROC score if I do not feel the accuracy score truly evaluates the models. I believe a 70% score would be good in this instance too.

Project Outline

I will start by examining the Starbucks data at a high level to get a feel for the data. This will help me understand the fields in each of the tables. I will also be able to see how each field is distributed. This will help me see if some demographic or one of the offers are barely represented and thus impossible to perform reliable analysis upon. This phase of the project will be a bit unstructured and free flowing as I go down different avenues trying to get a good feel for the data. Not all the analysis done here will be used later in the project.

I will use graphs to visualise the data. By using visual representations, it may also be easier to identify unusual outliers or data inconsistencies that may require cleaning. After cleaning and normalising the data, if necessary, it will be ready for the next phase. The thorough examination above will provide a good basis for identifying the fields and features will be most appropriate for the neural network models.

The data will then be split up into 3 testing, training and validation data. The training data will be the biggest. This will be used to train the neural network models. The validation sets will test the models during training. The testing data will then be used to see how the models perform at sending specific adverts to various customers in order to optimise the chance an offer will be completed.

The benchmark FNN model and the RNN model will both have the same testing, training and validation data so I will be able to compare how they compare to each other once the testing data is passed through the models. These will provide the final results of the project.

I will then discuss how I think the networks performed and how they can be improved upon. The results may provide insights on how the initial data analysis could be improved upon.

References

[1] Jin-A Choi, Kiho Lim, “Identifying machine learning techniques for classification of target advertising” in ICT Express (2020), pp. 175-180

[2] Chen Y., Kapralov M., Canny J., Pavlov D.Y., “Factor modelling for advertisement targeting” in Advances in Neural Information Processing Systems (2009), pp. 324-332