Machine Learning Engineer Nanodegree
Capstone Proposal

Sharvari Dhote

February 29, 2020

Domain Background

The businesses want to target right customers throughout the customer journey which helps to reduce the marketing budget, customer satisfaction as well increase the profits. Machine learning (ML) and artificial intelligence (AI) technologies are helping automate the digital marketing with real time decision making such as selecting best messaging platform, best timing, and the best offering to the individual customer by integrating data from different platforms [1, 3, 2, 4] as seen in figure 1.

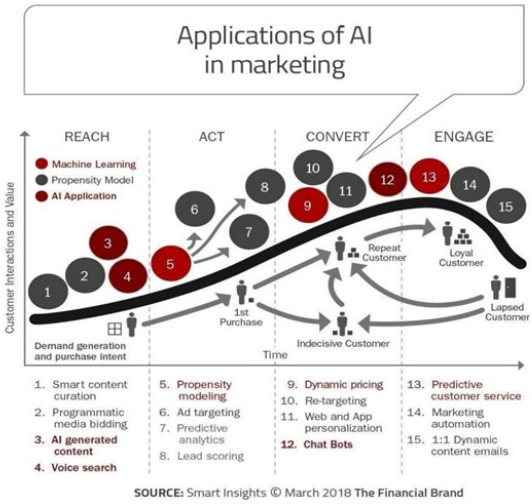

Figure 1: Application of AI in digital marketing [1].

Udacity's Starbucks Capstone challenge project data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Machine learning technique will be used in order to develop effective marketing strategy for the Starbucks business using the available data.

Problem Statement

Starbucks sends out an offer to users of the mobile app every few days. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). The given dataset contains demographic data a record of transaction and details about offer information. Before developing a predictive

model, exploratory data analysis and customer segmentation will be carried out which will help Starbucks business in better targeted marketing.
It is important to know your customers in order to increase business. First, customer segmentation using unsupervised machine learning technique K-means clustering will be carried out to understand current customer characteristics and demographics. Second the best performing supervised machine learning model will be build by comparing random forest and boosting ensemble models performance based on the defined metrics to predict if customer will respond to an offer or not for better targeted marketing.

Dataset and Inputs

Dataset : There are three .json files. Details are discussed below:

portfolio.json -(10,6) - information about offer type and duration

- $id(string)$ - offer id (10 offer sent - 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 (string -dict) (web, email,mobile, social)

profile.json -(17000,5) - demographic data for each customer - few missing values

- $age(int)$ - age of the customer
- $became\_member\_on(int)$ - date when customer created an app account
- $gender(str)$ - gender of the customer (note : 'O' for other , M or F)
- $id(str)$ - customer id (person - from transcript file)
- $income(float)$ - customer's income

transcript.json (306648,4) - with no missing values - records for transactions, offer types : received, viewed, completed.

- $event(str)$ - record description (ie transaction, offer received, offer viewed, etc.)
- $person(str)$ - customer id (id form profile file)
- $time(int)$ - time in hours since start of test. The data begins at time t=0
- $value - (dictofstrings)$ - either an offer id or transaction amount

Input : First cleaning, variable encoding, scaling and imputation etc. will be carried out to prepare input for the customer segmentation analysis and developing the predictive model. Features may be engineered or dropped depending upon the data analysis and segmentation study. A labeled dataset will be prepared for developing supervised learning model. The prepared dataset will be divided randomly in to train and test dataset. As mentioned in the Udacity Starbucks Capstone description, there will be a data cleaning and preparation challenge listed below when preparing input.

- Not all users receive the same offer
- Different validity period the offer type and informational offer to influence customer.
- Customer might make a purchase through the app without having received an offer or seen an offer
- A user can receive an offer, never actually view the offer, and still complete the offer

Solution Statement

Customer Segmentation Analysis - Unsupervised learning technique K-means algorithm will be used to cluster customers into groups [5, 6]. We may not need feature reduction, however, Principal Component Analysis(PCA) may be applied for the feature extraction.
Offer Prediction Model - Feature Engineering will be very important since we have to create a labeled dataset for developing the predictive model. A target output will be binary/multi-class will be decided based on the segmentation analysis and train and test dataset will be created. Linear regression, Random Forest, Decision Tree, Adaptive Boosting, and Gradient boosting model will be used for training the model [7].

Benchmark Model

Benchmark model will be linear regressor in the absence of business objective from Starbuck. In order develop a best performing model, grid search will be used to compare performance of all trained models [7].

Evalution Metrics

Following aetrics will be used for evaluating the model performance.

- Accuracy - The ratio of correctly predicted examples by the total examples. It shows often is the classifier correct. Accuracy may not be right metric always.
- Receiver operating characteristic (ROC) curve - more visual way to measure the performance of a binary classifier is . It is created by plotting the true positive rate (TPR) (or recall) against the false positive rate (FPR),
- AUC: relation between true positive rate and false positive rate - AUC stands for Area under the ROC Curve. It provides an aggregate measure of performance across all possible classification thresholds.
- Confusion Matrix - table 1 showing calculated correct predictions and types of incorrect predictions.

Project Design

Following are the project design steps [8, 9, 10]

- Data Exploration : First step is to know the data. Matplotlib/Seaborn/plotly libraries will be used for the exploratory data analysis (EDA). Study the data statistic, count missing values, plot different variables and understand and explore the data.

		Predicted class	
		Positive - Class 1	Negative - Class 2
Actual Value	Positive - Class 1	True Positive (TP) (Right)	False Negative (FN) (Wrong)
	Negative - Class 2	False positive (FP) (Wrong)	True Negative (TN) (Right)

Table 1: Confusion Matrix

- Model pre-processing and cleaning : Basic cleaning steps such as filling missing, checking outliers, dropping unimportant and duplicate columns. Encoding categorical variables. scaling, and feature engineering. Sklearn, numpy and pandas libraries will be used.
- Next, customer segmentation will be carried out to group the demographic and purchase data.
- Train and Test data : Labeled dataset will be prepared and split randomly to train and test dataset.
- Model training and implementation : Five different models will be implemented and based on the AUC value and other metrics, best model will be selected.
- Model validation : Test data will be used to validate the model performance with defined metrics.
- Model selection and optimization : Selected model will be tuned using a grid search.

References

[1] Application of AI in Marketing

[2] What Are The Technological Advancements In Marketing?

[3] How AI is Changing Digital Marketing

[4] Here's What You Need to Know About Propensity Modeling

[5] A gentle introduction to Customer segmentation

[6] The Most Important Data Science Tool for Market and Customer Segmentation

[7] Credit Card Fraud Imbalanced Dataset: Review, Metrics, Resampling techniques, Data Exploration and Comparing Machine Learning Models

[8] 11 Important Analytical Steps for your Data Science project

[9] A Data Science Workflow

[10] Comprehensive Guide to Data Exploration