# Udacity Machine Learning Engineer Nanodegree Capstone Project

Bikky Kumar

1. **Project Definition**
   1. Project Overview

The final project completed as a part of the Udacity’s ML ND program comes from Starbucks. This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. 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). Some users might not receive any offer during certain weeks.

Not all users receive the same offer. The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

Every offer has a validity period before the offer expires. For instance, a BOGO offer might be valid for only 5 days. It can be seen in the data set that informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, it is safe to assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.

The dataset for the problem consists of transactional data showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.

Someone using the app might make a purchase through the app without having received an offer or seen an offer.

The data for this project, as provided by Starbucks through Udacity, is contained in three files:

* portfolio.json – containing offer ids and meta data (reward, duration etc.) about each offer
* profile.json – demographic data for each customer
* transcript.json – records of transactions, offers received, offers viewed, and offers completed
  1. Problem Statement

The problem in the current project is divided into two parts – unsupervised machine learning task and supervised machine learning task.

* The unsupervised task is dividing data into various clusters and finding out offer responses based on these clusters. Specifically, efforts will be made to find out which cluster group best responds to which type of offers. These clusters will then be deconstructed into age and income groups in order to allow Starbucks to send targeted offers to people belonging to certain age, gender and income category.
* The goal of supervised machine learning task is to predict whether a customer will complete an offer sent by Starbucks. This will be a classification problem, wherein the features will be various demographics data (age, gender, income), offer types (BOGO, informational etc.) and personal spending habits. This model will help Starbucks predict the potential revenues generated through promotional offers in different regions.
  1. Metrics

The solution to this problem warrants two types of metrices – one for unsupervised task and other for supervised task.

The unsupervised task is related to finding optimum number of clusters of data. The optimum number of clusters will be decided based on elbow method. The elbow methods plot inertia score (within-cluster-sum-of-squares) against number of clusters. Optimum number of clusters are chosen based on lowering of inertia score.

Inertia

The k-means algorithm used for clustering, divides a set of N samples X into K disjoint clusters C, each described by the mean of the samples μj in the cluster. The means are called cluster centroids. The algorithm aims to choose centroids that minimize the following term known as inertia or within cluster Sum of Squares:

The supervised ML task uses a classification algorithm. The prediction from this algorithm will be a confusion matrix. The matrix will give us the values for True positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These values are based on predicted label and true label. Mainly two matrices will be used to evaluate the supervised part of the problem.

Accuracy (fraction of correct predictions)

Recall (fraction of all positive instances which classifier accurately predicts as positive)

1. **Analysis**
   1. Data Exploration and Visualization

The dataset for the project is given in three json files.

Portfolio

The portfolio file contains information about various offers sent by Starbucks. The file contains 10 rows and 11 columns. Columns are – offer id, offer type (BOGO, informational and discount), difficulty (minimum amount required to be spent), reward, duration, and channels through which offers are sent (email, mobile, social). A new field ‘name’ was created which contains all the information about an offer.

**A picture containing drawing

Description automatically generated**

For instance, the offer bogo\_10\_10\_5

represents a BOGO offer with difficulty of

10, reward of 10 and duration of 5.

Profile

The profile file contains information for 17000 persons – gender, age, income, person id, and data when the person became member of Starbucks rewards program. After exploring the data, it was observed 2175 persons with age of 118. These profiles also had missing income and gender information. These people were removed from the main analysis and analyzed separately in a special dataset. The main profile had people with genders noted as male, female and others. The distribution of data is as follows:

A picture containing screenshot, drawing

Description automatically generated

It can be observed that while most customers are males, the average income of females are higher than males. The other genders represent only a small percentage of the dataset.

Transcript

The transcript file contains information related to transactions – event, customer id, time, and value. The event is one of offer received, offer viewed, offer completed or transaction. The value field is a dictionary containing either offer id or transaction amount related to purchases made. In total the data contains 306534 event records with no missing values.

**A picture containing drawing

Description automatically generated**

The diagram on the right depicts the count of

various event types recorded by the app

A picture containing drawing

Description automatically generatedThe distribution of offers sent to customers

show that various offer types were equally

distributed among customers

A picture containing drawing

Description automatically generated

The average spending, which is an important feature for further processing of the data, shows bimodal log-normal distribution.

All the data files were combined into a single file for further processing.

1. **Methodology**
   1. Data Preprocessing

The main preprocessing done on data is as follows:

* Channels feature in portfolio dataset is one-hot-encoded
* Became\_member\_on feature in profile dataset is parsed into date format
* The transcript dataset contains a field value, which contains a dictionary object of offer id and transaction amount. This field is parsed to create two separate field offer id and amount.
* All the datasets are combined to create a single dataset using Pandas’ merge functionality for further analysis
* One-hot-encoded member’s gender, joining year and joining month
* Additional features are created using original features – one-hot-encoding of various events, average spending per customer, total spending per customer, transaction count per customer, invalid flag (customers who completed offer without having seen the offer), rate of offer received, viewed and completed for each offer type, rate of completed transactions.
* The intermediate preprocessing steps included Standardization of data and calculation of Principal Components
* For supervised machine learning task, target is calculated by assigning 0 to customer who did not complete any offer while 1 was assigned to customers who completed at least one offer
  1. Implementation

Unsupervised ML part

The unsupervised ML part of the problem comprised of dividing data into meaningful clusters. The k-means algorithm was used to divide data into clusters. The data was divided into two – main profile (with complete data fields) and special profile (with age, income and gender data missing). The special profile represented only a small percentage of total data. In this report, the details of only main profile is reported. However, similar analysis was done on special profile as well.

A close up of a colorful background

Description automatically generated

The Elbow plot method was used to find optimum number of clusters. Although it was not completely clear from the plot, it was decided to divide data into 10 clusters.

A picture containing screenshot, drawing

Description automatically generatedThe data was divided into 10 clusters and normalized cluster centroids are plotted for various parameters. The figure on the left shows cluster centroids for completion rate of various offers sent by Starbucks – BOGO, Discount, Informational and Total

A screenshot of a cell phone

Description automatically generated

The data shows normalized centroid values for various types of discount offers

A screenshot of a cell phone

Description automatically generated

Normalized centroid values for various BOGO offers

A picture containing screenshot, drawing

Description automatically generated

Normalized centroid values for various Informational offers

A picture containing room

Description automatically generated

Comparison of centroids for some important features for various clusters

A close up of a logo

Description automatically generated

Gender distribution for various clusters

Supervised ML part

The supervised ML part of the problem comprised of creation of a classification model to predict which customers are likely to complete an offer sent by Starbucks. A baseline model was generated using Logistic Regression. The confusion matrix generated from the model is as follows:

A screenshot of a cell phone

Description automatically generated

The accuracy of this logistic regression model is 85%. Label 0 means the customer is not likely to complete an offer, while label 1 depicts customers who are likely to complete an offer.

* 1. Refinement

After an initial baseline model, an advanced model Random Forest was used. One of the motivations of using this model was to check for feature importance. The accuracy of this model was 86%, not much different than the baseline model.

A close up of a logo

Description automatically generated

The relative feature importance was plotted using Random Forest classifier. Most of the features related to member month were not very significant. It was hypothesised that customers who became members during holiday month like December would be more likely to make purchases. However, it was not the case. These parameters were dropped for further analysis.

The next model used was Support Vector Machines classifier. The accuracy of the model was again 86%, similar to baseline model.

It was important to understand if accuracy was the right evaluation metrics for the problem at hand. The accuracy metrics maximized the prediction of TP and TN. However, in this case it is not that important to predict TN. It does not cost Starbucks a lot of money to send offers to customers, who are wrongly predicted to complete the offers. However, it is really important to target customers who are likely to make purchases. Considering this fact, the better metric to optimize in this case would be recall, which maximized TP rate.

Sklearn’s Grid Search functionality was used to optimize recall score using SVM. The confusion matrix from optimized solution is as follows:

A screenshot of a cell phone

Description automatically generated

The accuracy for this model was 75%. However, this model was able to maximize TP rate at the expense of FP. The recall was 98%. This is a better model for present purposes.

1. **Results**

Unsupervised Model

The unsupervised model was able to segment data into various clusters. Some of the salient observations from the data grouping are as follows:

* Clusters with highest rate of overall offer completions are – 1,3,4,7,8, and 9
* Clusters with highest BOGO completion rate – 1,3,7,8, and 9
* Clusters with highest Discount completion rate – 0,4,6, and 7
* Clusters with highest Informational viewing rate – 5

Starbucks can send offers to targeted cluster groups in order to increase its chances of success.

In terms of Discount offers:

* Cluster 0 was very responsive to offer with difficulty level 7
* Clusters 4 and 6 was very responsive to offer with difficulty level 10
* Clusters 7 was very responsive to offer with difficulty level 2

For BOGO offers:

* Clusters 1 and 9 were very responsive to offer with difficulty level 5
* Clusters 3 and 8 were very responsive to offer with difficulty level 10

These clusters can then be described in terms of demographics data for easy targeting by Starbucks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Clusters** | Offer completion rate | Age | Income | Avg\_spending | Transaction Count |
| 0 | 0.37 | 52.9 | 63325.6 | 12.6 | 10.0 |
| 1 | 0.41 | 54.3 | 66261.9 | 15.4 | 9.2 |
| 2 | 0.00 | 51.5 | 58133.1 | 7.2 | 5.0 |
| 3 | 0.52 | 55.6 | 69760.1 | 19.6 | 9.8 |
| 4 | 0.55 | 55.5 | 68041.7 | 18.4 | 9.5 |
| 5 | 0.00 | 53.0 | 59957.2 | 9.5 | 5.7 |
| 6 | 0.34 | 53.6 | 63800.3 | 13.8 | 9.9 |
| 7 | 0.65 | 56.7 | 70043.6 | 20.8 | 9.4 |
| 8 | 0.47 | 57.2 | 73492.6 | 21.3 | 8.8 |
| 9 | 0.55 | 55.7 | 68674.3 | 18.1 | 9.2 |

A few things can be noted:

* Clusters with higher offer completion rate have higher income and high age
* Income is highly correlated with average spending

Supervised Model

The final model for supervised part of the problem was a support vector machine classification model. The hyperparameters used for this model are as follows:

|  |  |
| --- | --- |
| C | 1.0 |
| Gamma | 0.5 |
| kernel | rbf |

As stated earlier, although the accuracy of this model was lower than baseline case, this model was more suitable for problem at hand, that is, predicting customers who are more likely to complete an offer based on demographics and offer details.

The classification report for this model is as follows:

**precision recall f1-score support**

0 0.74 0.12 0.21 1244

1 0.74 0.98 0.85 3199

**accuracy**  0.74 4443

**macro avg** 0.74 0.55 0.53 4443

**weighted avg** 0.74 0.74 0.67 4443

The supervised ML part of the problem is relatively simple, with a binary classification problem. Most of the classifiers would work fine. The most important thing to consider is finding the right metric to optimize. As discussed, accuracy might not be the best parameter to optimize.

1. **Conclusion**
   1. Reflection

Overall the Starbucks problem is a classic case of customer segmentation. Although the problem is specific to Starbucks in the present case, it can be easily expanded to other organizations. Similar organizations use mobile app to track customer behavior. This information can be used to send targeted offers. It serves both ways – organizations can make more money and customers can get what they like most. As is the case in most machine learning tasks, the model and prediction get better with more data. This problem can be expanded to include more of Starbucks products.

One of the areas where this problem was challenging was during clustering process. Although the data was separated into clusters, there was considerable overlap between clusters. This problem can be overcome by including more features.

* 1. Improvement

Although the problem was able to be separated into clusters, there is a significant overlap between different clusters. The main parameters affecting the clustering process are age and income. More parameters are needed to separate the clusters in a better manner. Some of these parameters could be related to geographical information and information related to socio-economic conditions of that geographical locations. The effect of age and income would then be moderated according to the local conditions. This way Starbucks will be able to send offers which are also targeted according to person’s location.

The supervised task can be improved by including more features. The features related to offer viewing by the customer was not included in the modeling. If this information is included, it could lead to a better prediction. It does have some caveats, as this information is won’t be available to Starbucks at the time of offer sending.