2018

Acquire Valued Shoppers

PREDICTC WHICH SHOPPERS WILL BECOME REPEAT BUYIERS

MUDASSIR ALI SYED

Table of Contents

Abstract

Businesses today often face a challenge of retaining customers, and converting them to be regular shoppers. Businesses run huge sales with discount coupons/offers to attract customers and hope to convert them as regular buyers. This project deals with almost 350 million rows of completely anonymized transactional data from over 300,000 shoppers along with offer details and History of offers given to customers.
We will use the provided data to predict the customers who will be repeat buyers.

Introduction

Background

Technology has created a huge shift in the way customers can buy products, to adapt to this change consumer brands are also changing their business models to acquire customers.

Consumer brands often offer discounts to attract new shoppers to buy their products. The most valuable customers are those who return after this initial incented purchase. With enough purchase history, it is possible to predict which shoppers, when presented an offer, will buy a new item.

LAER is a sales term used in any business

Land
Adopt
Expand
Renew

- **Land:** *"All sales and marketing activities required to land the first sale of a solution to a new customer."* When you land the customer, you've successfully convinced the prospect to become a new customer of yours.

- **Adopt:** *"All activities involved in making sure the customer is successfully adopting and expanding their use of the solution."* This is the step where you help the customer that just bought your product.

- **Expand:** *"All activities required to* cost-effectively help current customers expand their spending *as usage increases, including both cross-selling and upselling."* As you become more invested in the customer's outcomes, it becomes easier to tie your technology to other projects and initiatives, encouraging your customers to buy more products and services from you the supplier.

- **Renew:** *"All activities required to ensure the customer renews their contract(s*)." Convincing your customer to renew their relationship with you when it comes time to repurchase.

All the above aspects of sale are very important, but Renew is extremely critical for businesses to operate in long run.

Related to this is our project where store provides its customers with a discount offer and we will be predicting the possibility of customer returning to the store and buying the same item.

This will be a classification problem where a customer will repeat the purchase or will not be repeating the purchase.

We will be doing feature extraction and applying different machine learning models to derive AUC score.

About Data

The data set is obtained from Kaggle competition, which is publicly available in https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data

Dataset has four relational files:

transactions.csv - contains transaction history for all customers for a period of at least 1 year prior to their offered incentive
trainHistory.csv - contains the incentive offered to each customer and information about the behavioral response to the offer
testHistory.csv - contains the incentive offered to each customer but does not include their response (you are predicting the repeater column for each id in this file)
offers.csv - contains information about the offers

Transaction data has almost 350 million rows of completely anonymized data from over 300,000 shoppers along with offer information and History of offers given to customers.

The size of the dataset if 3GB.

Due to size constraints, we will choose 10000 customers out of 16000 who were given a voucher/offer.

Following are the features of the datasets

All the fields are anonymized and categorized to protect customer and sales information. The specific meanings of the fields will not be provided (so don't bother asking). Part of the challenge of this competition is learning the taxonomy of items in a data-driven way.

Transactions

1. id - A unique id representing a customer
2. chain - An integer representing a store chain
3. dept - An aggregate grouping of the Category (e.g. water)
4. category - The product category (e.g. sparkling water)
5. company - An id of the company that sells the item
6. brand - An id of the brand to which the item belongs
7. date - The date of purchase
8. productsize - The amount of the product purchase (e.g. 16 oz of water)
9. productmeasure - The units of the product purchase (e.g. ounces)
10. purchasequantity - The number of units purchased
11. purchaseamount - The dollar amount of the purchase

History

1. id - A unique id representing a customer
2. chain - An integer representing a store chain
3. offer - An id representing a certain offer
4. market - An id representing a geographical region
5. repeattrips - The number of times the customer made a repeat purchase
6. repeater - A boolean, equal to repeattrips > 0
7. offerdate - The date a customer received the offer

Offers

1. offer - An id representing a certain offer
2. category - The product category
3. quantity - The number of units one must purchase to get the discount
4. company - An id of the company that sells the item
5. offervalue - The dollar value of the offer
6. brand - An id of the brand to which the item belongs

Data Exploration

We can see the distribution of how many customers returned when provided an offer/discount, based on given dataset.

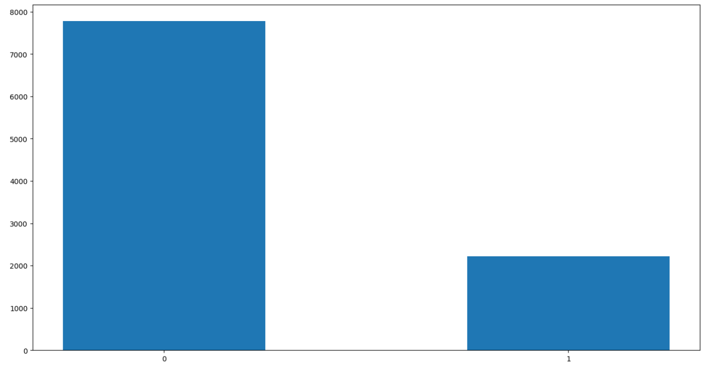

Figure 1.

Figure 1 shows, the histogram of returning customer distribution, we can see close to 25% of customers seem to have repeated purchase when given a discount.

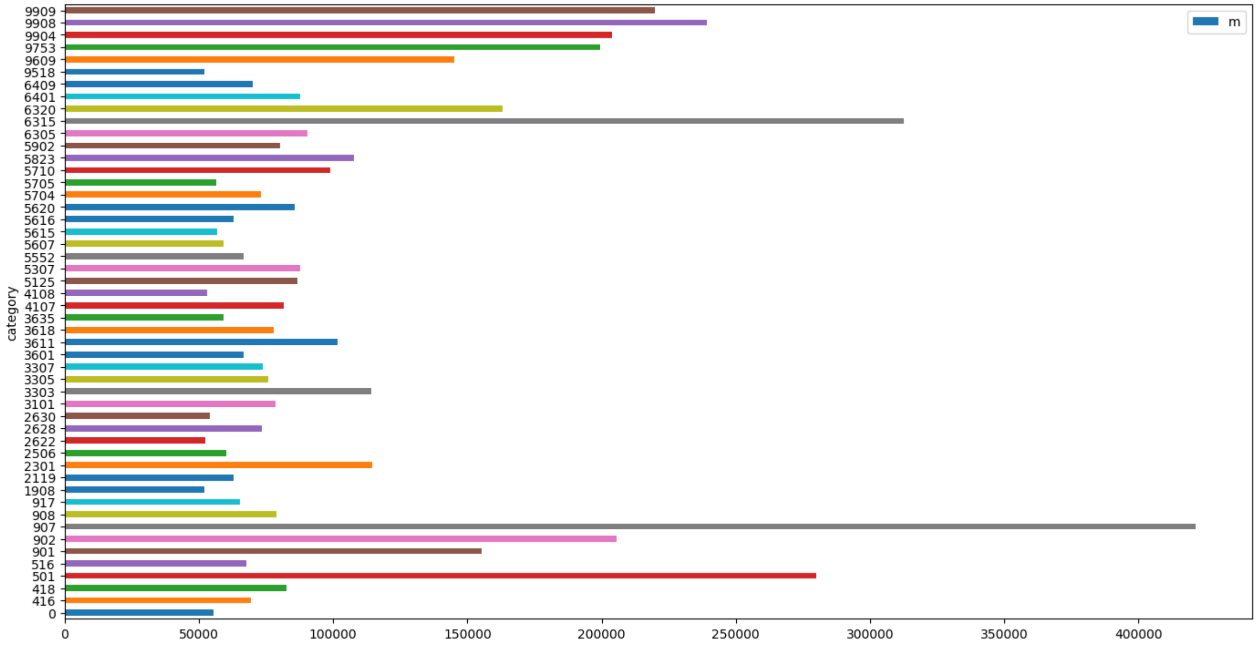

Figure 2.

Figure 2 shows, a bar chart of top 50 categories with high transactions.

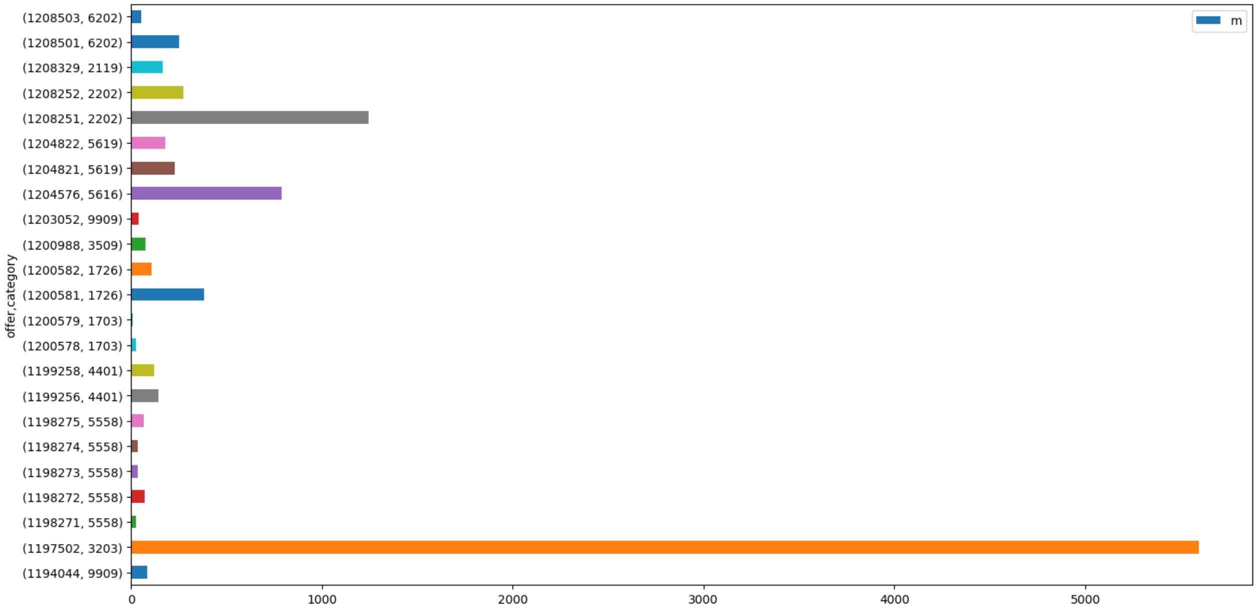

Figure 3.

Figure 3 shows, a bar chart with number of customers to which each offer has been given.

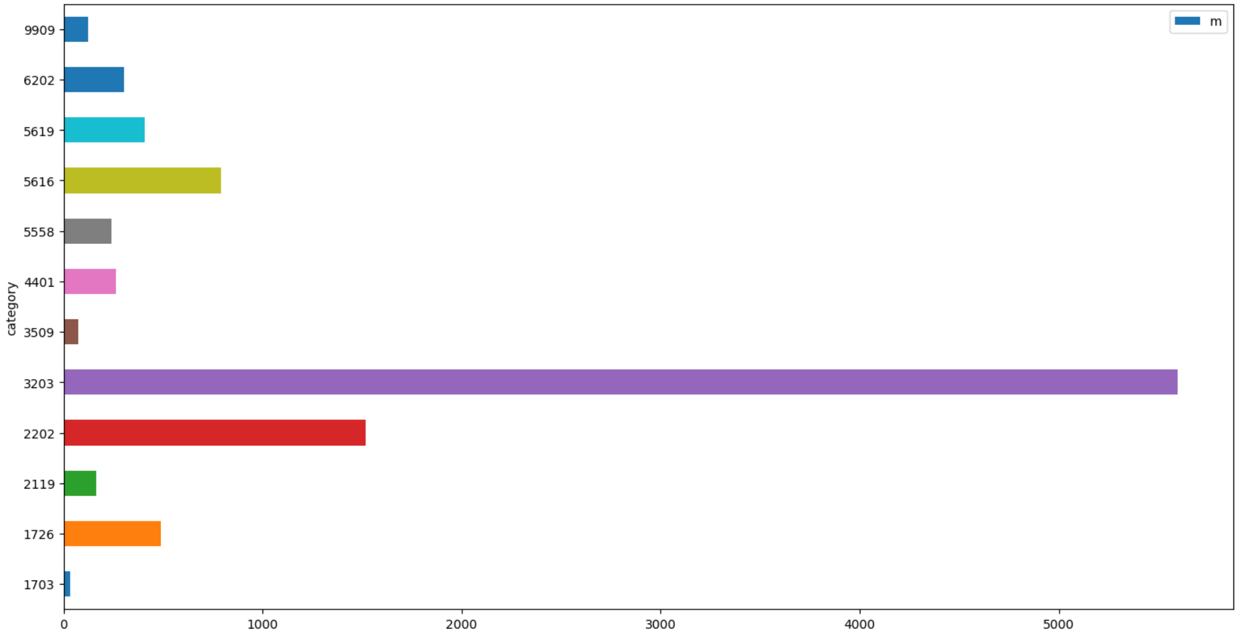

Figure 4.

Figure 4 shows, a bar chart number of offers given for each of the categories.

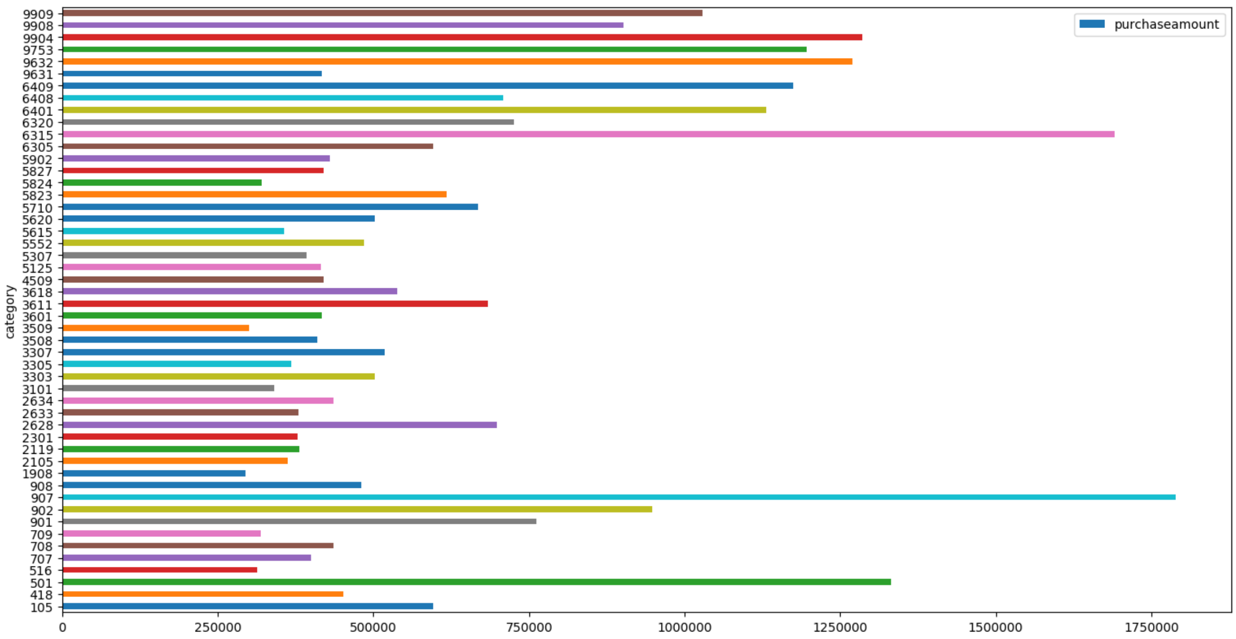

Figure 5.

Figure 5 shows, a bar chart shows top 50 popular categories based on purchase.

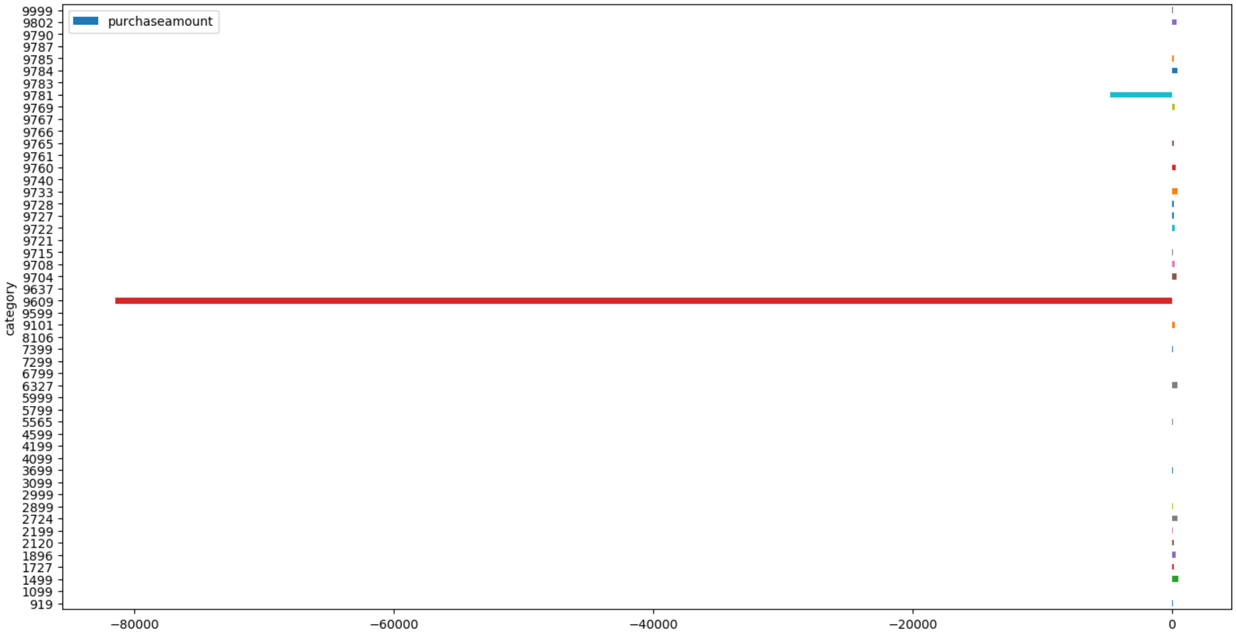

Figure 6.

Figure 6 shows, a bar chart shows 50 unpopular categories based on purchase amount.

Predictive Task

Our predictive task is to forecast if the customers offered a discount coupon will return and become a repeat buyer.

Baseline

We establish base lines with given train history of customer who were offered discount offers. Logistic regression with 100-fold cross validation will result in AUC score of **0.5148**.

Validation

We will be validating our model and tuning the hyperparameters using a 100-fold cross validation. We randomly shuffled our data set and used 1/3 of the data points as test set. We partitioned the remaining data set into 100 splits. In each of the 100 iterations we trained our model on 99 splits and calculated the AUC score on the remaining split. Based on the average of 100 AUC, we tuned our hyper parameters.

Feature Extraction

Models

We run a model to predict the repeat buyers, we explored different types of parametric models based on the features mentioned above. Since this is a regression problem, we will choose a classifier that can efficiently train our model, while considering that we have 10000 customer data in our training set and around 12 features.

Evaluating the Model

Logistic Regression

Decision Tree

Random Forest

Results

Conclusion