PES UNIVERSITY, BANGALORE

Department of Computer Science and Engineering

Abstract

Santander Bank is one of North America's top retail banks by deposits and is wholly owned by Santander Group, serving more than 100 million customers in the United Kingdom, Latin America, and Europe. Santander Bank offers a lending hand to its customers through personalized product recommendations to support needs for a range of financial decisions.
This paper focuses on providing a recommendation system that will help the bank predict which products their existing customers will use in the following month.

Introduction

Under their current system, a small number of Santander's customers receive many recommendations while many others rarely see any resulting in an uneven customer experience. With a more effective recommendation system in place, Santander can better meet the individual needs of all customers and ensure their satisfaction no matter where they are in life.

The training data consists of nearly 1 million users with monthly historical user and product data between January 2015 and May 2016. User data consists of 24 predictors including the age and income of the users. Product data consists of boolean flags for all 24 products and indicates whether the user owned the product in the respective months. The goal is to predict which new products the 929,615 test users are most likely to buy in June 2016. A product is considered new if it is owned in June 2016 but not in May 2016. The next plot shows that most users in the test data set were already present in the first month of the train data and that a relatively large share of test users contains the first training information in July 2015. Nearly all test users contain monthly data between their first appearance in the train data and the end of the training period (May 2016).

A ranked list of the top seven most likely new products is expected for all users in the test data. The leaderboard score is calculated using the MAP@7 criterion. The total score is the mean of the scores for all users. When no new products are bought, the MAP score is always zero and new products are only added for about 3.51% of the users. This means that the public score is only calculated on about 9800 users and that the perfect score is close to 0.035.

The test data is split randomly between the public and private leaderboard using a 30–70% random split.

EVALUATION METRIC

The evaluation metric used is Mean Average Precision @ 7 (MAP@7):

$$MAP@7 = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{min(m,7)}\sum_{k=1}^{min(n,7)}P(k)$$

Where

|U| is the number of rows (users in two time points)

P(k) is the precision at cutoff k

n is the number of predicted products

m is the number of added products for the given user at that time point.

Related Work

1) Content-Based Filtering

Content-based recommender systems work with profiles of users that are created at the beginning. A profile has information about a user and his taste. In the recommendation process, the engine compares the items that were already positively rated by the user with the items he didn't rate and looks for similarities. Those items that are mostly similar to the positively rated ones, will be recommended to the user.
This approach is not possible to meet the needs of our problem statement since it is difficult to define the products of banking (e.g. Credit account, Savings account, Payroll account etc.) with its content.

2) Collaborative Filtering

The idea of collaborative filtering is in finding users in a community that share appreciation. If two users have the same or almost the same rated items in common, then they have similar tastes. Such users build a group or a so-called neighbourhood. A user gets recommendations to those items that
he/she hasn't rated before, but that was already positively rated by users in his/her neighbourhood.

3) Item-Based Approach

These approaches are most followed by various organizations to recommend products for their existing customers while they do have a cold start problem with new customers. A decent number of programmers tried collaborative filtering-based approaches by clustering users based on their preferences and recommending the products most preferred in the respective cluster. However, this type of approach doesn't consider the situation or context in which the user preferred that particular product.

Problem Faced With The Dataset

Model-based algorithms are widely used today, where dimension reduction and matrix factorization techniques belong to this group of algorithms.

DataSet is in Spanish, formatting and translating to English is a prerequisite before proceeding further.
The traditional content-based approach is not possible to meet the needs of our problem statement since it is difficult to define the products of banking (e.g. Credit account, Savings account, Payroll account etc.) with its content.
Collaborative filtering approaches are most followed by various organizations to recommend products for their existing customers while they do have a cold start problem with new customers. Many analysts have tried filtering-based approaches by clustering users based on their preferences and recommending the products most preferred in the respective cluster. However, this type of approach doesn't consider the situation or context in which the user preferred that particular product.

What is the specific problem we are going to solve?

The main aim is to be able to recommend financial products to customers based on their behaviour data, this would help a bank serve its customers even better while increasing its profits.

METHODOLOGY

1) Dataset

The datasets used in this project are taken from Kaggle and are available at the following link:

https://www.kaggle.com/c/santander-product-recommendation

The dataset contains information on customers and product data between January 2015 and May 2016. User data consists of 24 predictors including the age and income of the users.

2) Preprocessing data

Our dataset has approximately 13 million rows with 48 columns of data. Therefore with such a high volume of data, the number of null values and outliers, and duplicate values should be taken care of. We have found that 3 columns of data have a majority of null values and so have dropped the 3 entirely ('last_date_as_primary', 'spouse_index', 'address_type').

On further processing, the number of null values was still high. To take care of this, all these null values have been dropped from the data set using the .dropna() command.

DATA ANALYSIS

To understand the data, and the features that will be needed to play an important role in our model, we have performed Exploratory Data Analysis and Visualization to select the important features.

Figure 1 shows the trends in the variation of purchases made by new customers.

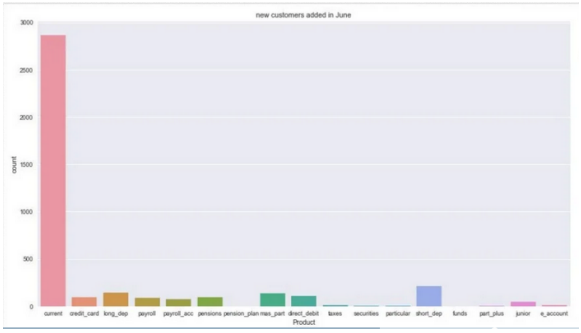

Fig 1

Figure 2 shows the variation in the age of the customers along with grouping by their genders which can help us to find any pattern based on their shopping trends based on their age.

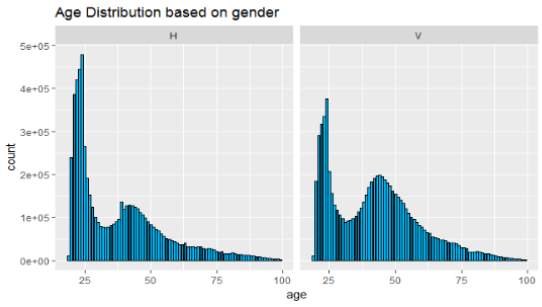

Fig 2

In figure 3, the number of customers vs the number of products bought is depicted on a per-month basis, showing us the variation in the trends.

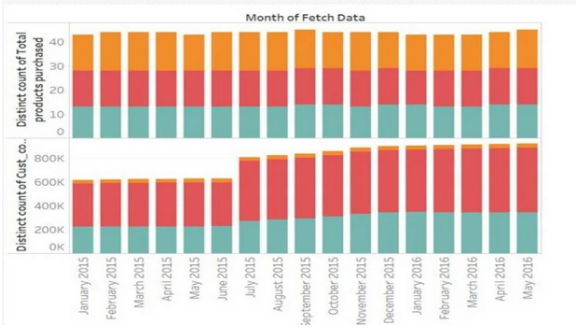

Fig 3

The correlation becomes an important metric to understand the data as it tells us which attributes are closely related to each other as depicted in figure 4. It can also help us later in dropping attributes to increase model performance.

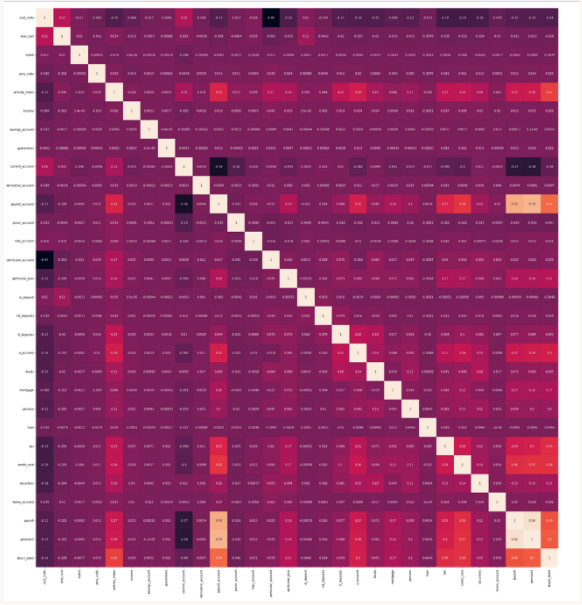

Fig 4

References

- https://dl.acm.org/doi/abs/10.1145/3404835.3464928
- https://www.kaggle.com/c/santander-product-recommendation
- https://towardsdatascience.com/recommendation-systems-explained-a42fc60591ed

By:-

1. Mithul Chander B - PES2UG20CS198
2. Praneeth Kumar L - PES2UG20CS251
3. Rehan Ganapathy - PES2UG20CS269

https://github.com/rehanganapathy/data-analytics-project