Project Proposal
CSCI-P556: Applied Machine Learning
Fall 2018

Parth Naik (naikpa)
Pranay Shah (shahprn)
Mayank Kumar Raunak (mraunak)

October 5, 2018

1 Problem Statement

Our project was a past Kaggle Competition posted by Santander (Santander Product Recommendation) in which the company provided data of customers. The data was accumulated by the bank over a period of 1.5 years. Using this data a prediction has to be made on what new products the customers are more likely to purchase. The data has monthly records of products a customer has such as "Credit Card", "Savings Account", etc. This project is particularly interesting to us because in any business venture to increase revenue, marketing plays an important role and if we can identify the customers' needs', map it to the products or services offered by the business, in our case, "Santander", and sell them to the customer, then this is a win-win situation for both the parties involved. This would lead to an increased customer satisfaction, and due to the idea of promoting services in accordance to the person's needs, the organization may also see an increased number of potential customers.

2 Data

1. Description of Data
2. Domain of Data : Bank Customer Data
3. Size of Data : 2.3GB
4. Samples : 13647309
5. Features : 48
6. Missing Values : Two features have a lot of missing values(Can be dropped)
7. Imbalanced : It is a regression problem
8. Categorical Features : 17
9. Numeric Features : 31
10. Binary : 0
11. Data Acquisition : Kaggle

3 Questions

- Pair the customers with the products by recommending additional personalized products for the next month.
- Predict the credit limit and loan to be given to a particular customer as well as a particular group of customers.
- Predict the product needed by a particular customer based on the locality.
- On the basis of product purchased by a customer, predict customer's age, gross house hold. income,locality, his resident country etc.
- Design an experiment to find if the noise data gives signal or it should be eliminated.

4 Evaluation Criteria

- Evaluation on basis of 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, and m is the number of added products for the given user at that time point. If m = 0, the precision is defined to be 0.

5 Timeline and Roles

- Data Analysis and Visualization : 10/13/2018 - 10/26/2018
- Building and selecting the models: 10/27/2018 - 11/15/2018
- Reporting results : 11/16/2018 - 11/20/2018

Team Members	Questions
All	1
Parth Naik	2
Pranay Shah	3
Mayank	4