Banking Product Recommendation System

Project Description

A Spanish bank is looking for a better recommendation system to improve its current one. Under their current system, a small number of customers receive many recommendations while many others rarely see any, this results in an uneven customer experience. Our task is to predict which products existing banking customers will use in the next month based on their past behavior and that of similar customers.

With a more effective prediction/recommendation system in place, the Spanish bank will be able to better understand the needs of their customers and provide better product offerings to them thereby enhancing the customer experience. This in turn should improve their sales and marketing performance significantly as well.

The topic for this project is inspired by the competition that is conducted by Banco Santander through Kaggle. The overview of the competition can be found in the following link.

Research Questions

1. Is there a correlation between the customer demographic data and product(s) that the customer is actively using?
2. What machine learning models can handle time-series for multiple predictor variables (multivariate time series data)? The data is monthly for approximately 18 months and for a range of financial, socio-economic predictors and bank products.
3. Given there are 25 predictor variables, what are the variables necessary to make an accurate prediction?
4. Given there are 24 response variables, is it necessary to produce separate models for each response variable? Or are there particular machine learning models that can make predictions for multiple response variables?
5. Do we want to make the model more generic and able to handle new customers (with limited data availability), not just existing customers?
6. Can the final model and its prediction be used by the sales/marketing team to improve the efforts for a more targeted marketing strategy?

Methodology

For our project, we plan to use the dataset that is shared by Banco Santander as part of the competition that they conducted six years ago for a similar topic, *Bank Products Recommendation System*. The train and test dataset can be found in the following link. Since the competition was concluded six years ago, we have no access to the true value of the test dataset hence we will only use the training dataset and split it to training and testing dataset.

We plan to perform the Exploratory Data Analysis to understand the distribution of the data, as well as the correlation between variables. We also plan to perform Feature Engineering to select only variables that have predictive power for the product recommendations system.

The following machine learning models can be used to investigate the research questions posed above:

ARIMA

It would be useful to use time-series analysis / techniques on the data to see if there are any seasonality effects that can be taken into account when offering customers certain products. Some techniques that can be looked at are the Autoregressive Integrated Moving Average (ARIMA) or Seasonal Autoregressive Integrated Moving Average (SARIMA).

Classification Models

Various regression methods such as multiple linear regression, stepwise regression and principal component analysis allow us to observe the importance of each variable and selecting the most important ones. Making use of only the important variables or components of the important variables would help us create a more accurate model. Classification methods such as the KNN, PCA-KNN and decision trees would allow us to see the ranges of the significant variables that help to make an accurate prediction.

Other Models

Other than the models mentioned above, there are a number of models that would be helpful in making predictions of a customer's future products. These models are powerful at making predictions but may not be helpful in understanding the importance of predictor variables. Some models that can be explored from a purely prediction perspective are: Random Forest, Association Rule Learning, Neural Networks and other ensemble methods.

Cross validation will be conducted to ensure the validity and accuracy of selected models.

We believe it would be helpful to look at the cost of false positives or false negatives. In the scenarios of banking products, these would come in the form of the cost of improving experience by spending more time with the customer in order to sell a product as well as the cost of not selling the product to an interested customer. These would allow us to further understand the profitability of each model.

Our Team

Chen, Lulu - Graduated in Actuarial Science from Simon Fraser University in Canada. Now work full-time in Prudential Hong Kong as an actuary.

Chua, Christian – Graduated in Business Management from Singapore Management University. Currently working in PwC on data management projects.

Wallace, Bradley – Graduated as a civil engineer from the University of Adelaide. Now working and studying Korean part-time, alongside the Master's degree, while living in South Korea.

Wiratama, Ardy - Works as a Senior Industry Process Consultant for Dassault Systemes and serves the energy and resources industry for over 10 years. He is based in Perth, Australia.

Zhang, Yichi – Graduated as a computer engineering student from National University of Singapore. Now working in Accenture as a consultant specializing in wealth management and advisory solutions.