Project Milestone Two – Individual Project

Logan Quandt

College of Science and Technology, Bellevue University

DSC630 – Predictive Analytics

Dr. Fadi Alsaleem

June 19th, 2022

**Introduction:**

Many people might not think about how Banks make money and consider that their deposits are considered liabilities by the bank. While banks can make money by having large deposit accounts and lending the money overnight to other financial institutions to meet their minimum reserve requirements, the primary revenue generator for banks is loaning out money and receiving interest payments. It’s for this reason that banks spend tons of money and time marketing to not only new customers but existing deposit account holders to cross-sell them on new loan products.

Having previously worked as a bank teller doing the cross-selling I was interested in the possibility of using predictive analytics to better focus marketing efforts for cross-selling deposit holders and increase the conversion rate for these campaigns which tends to be in the single digits. In this effort I will be using segmentation and classification to determine whether a customer is a good target for a personal loan marketing campaign.

**Dataset:**

For this project I was able to find a dataset from Kaggle titled “Bank Loan Modeling.” The dataset simulates information on 5000 bank customers from a previous marketing campaign and has 14 features including:

* ID: Customer ID
* Age: Customer's age in completed years
* Experience: # of years of professional experience
* Income: Annual income of the customer ($000)
* ZIP Code: Home Address ZIP code.
* Family: Family size of the customer
* CCAvg: Avg. spending on credit cards per month ($000)
* Education: Education Level.
  + 1: Undergrad
  + 2: Graduate
  + 3: Advanced/Professional
* Mortgage: Value of house mortgage if any. ($000)
* Personal Loan: Did this customer accept the personal loan offered in the last campaign?
* Securities Account: Does the customer have a securities account with the bank?
* CD Account: Does the customer have a certificate of deposit (CD) account with the bank
* Online: Does the customer use internet banking facilities?
* CreditCard: Does the customer use a credit card issued by the Bank?

With this dataset, the goal of the bank’s is to classify customers and predict the likelihood that a bank customer would accept a personal loan offer if included in a marketing campaign to better focus marketing efforts.

**Model Selection:**

For this project, I plan on using multiple different models including Logistic Regression, KNN (K-Nearest Neighbor), Decision Tree Classifier and Random Forest Classifiers to predict whether a customer will accept a personal loan and to use EDA to identify what segment of customers should be targeted in future campaigns. The target variable will be the “Personal Loan” Feature (0 or 1) and I will initially plan to use all the other features (besides ID) as predictors though this is subject to change after I explore the data and test out the models.

**Model Evaluation:**

As this is a classification problem at heart, I will be evaluated using the following to evaluate each model.

* Accuracy Score – To evaluate the number of correctly predicted records. This may not be as helpful as my dataset is imbalanced though I plan to run it on each model.
* Precision – To evaluate how accurate the models are when it comes to predicting positive values
* F1 Score – The mean of Precision and Recall
* Recall – To evaluate the total number of positives predicted by my model
* Confusion matrixes – This will allow me to evaluate the number of true positives, false positives, true negatives and false negatives.
* ROC AUC curves – To evaluate the performance of my models at various thresholds

**What do you hope to learn?**

For many previous course projects, I have done regression problems. By taking on this bank customer classification problem I am hoping to expand my knowledge of classification models.

**Risks/Contingencies:**

The dataset is small with only 5000 records and is made up of simulated information that is imbalanced. I will have to keep on eye on whether I need to tune parameters or use random sampling (over/under) to help the models accurately distinguish between the majority and minority classes. It is also very possible that the simulated data may perform poorly with some models, so I have chosen to use multiple to compare and

If I am unable to use this data for my project, I have identified a secondary dataset containing information on credit card customers that can also be used for making predictions on bank customers and whether they would be a good candidate for a credit card marketing campaign.

**References:**

<https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling>

[Evaluation Metrics for Classification Problems with Implementation in Python | by Venu Gopal Kadamba | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-problems-with-implementation-in-python-a20193b4f2c3)