LaTeX STAT451 Project Report: Online Banking

Claire Westenberger
cwestenberge@wisc.edu

Jessica Ostroff
jaostroff@wisc.edu

Manav Kalathil
kalathil@wisc.edu

Abstract

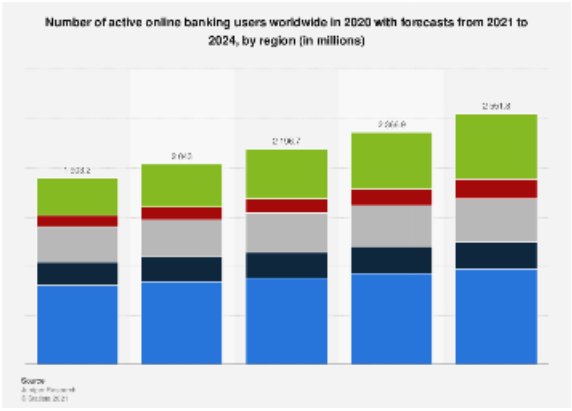
With the internet growing day by day, companies need to keep up to date on the wants of their clientele. In particular, banks need to be aware how their customers are utilizing online banking services. A prediction where a customer's background, including education level, family status, and other key factors are taken into consideration in order to better forecast their banking activity can be very powerful. The benefit of having this model is that it will help optimize a bank's account holder acquisition process, as the model can be used to predict whether or not a customer will be interested in using online banking services.

The motive behind this research is to take a customer's background and use that information to predict their use of online banking. Online banking services have been rapidly growing and it is crucial to banks, whether small or large, to optimize their customer satisfaction in order to stand out [4]. As banking becomes more and more digitized, this research can can be used predict whether the given customer will utilize internet banking services, and give an inside look on what demographics are correlated with this decision. This research can skyrocket a bank into success because they will have a greater understanding of the factors that affect online banking for each given customer based on their activity.

Through this study, banks will be able to understand what factors are most related to successful online banking and create and modify their products accordingly. This process will make the customer happy. With a happy customer there is deeper loyalty to the bank and more money being put into the bank.

1. Introduction

In today's fast paced world, banks are realizing that it is no longer sufficient to offer customers a solely brick and mortar experience. With technological leaps being made at a break neck pace, using online banking has never been easier or more accessible. As of 2021, approximately 2.04 billion people use online banking world wide [2], with 64.6 percent of US citizens using online banking [1]. This represents a massive market for banks. By adding online banking services to their suite of services, they can attract new and younger customers.



Online banking users worldwide in 2020 with forecasts to 2024, by region [2]

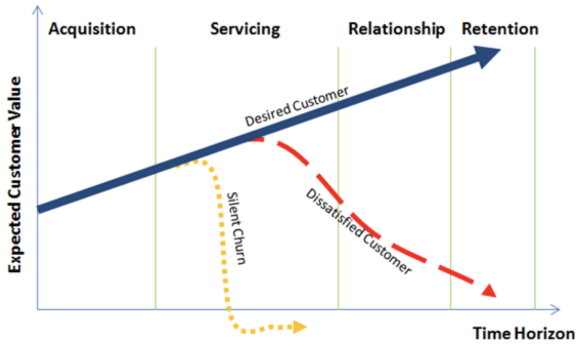
Additionally, online banking will expand banking usage as it can reach a larger amount of people through smartphones and devices with access to internet. It allows for users to interact with and conduct their financial transactions at their own convenience. Having the ability to pay bills, transfer money, as well as access specific records of one's checking account transactions from the comfort of one's home makes for easier financial work. Online banking eliminates the burden of physically going to a bank, which many may be unable to do, due to lack of transportation, or being unable to go to the bank during its hours of operation. Additionally, due to Covid-19, many people are not comfortable with heading into a packed bank.

Prior to the advent of mobile banking, individuals would have to miss work in order to make it to the bank, or additionally, would not complete their financial tasks in time due to limited access. Current users of this online service chose to use this mobile version as it allows for "on-the-go" interaction at all times all over the globe [7]. Online banks usually provide higher interest rates compared to it's in-person banking ways for the same type of savings account, and many offer interest on checking accounts, which is rarely offered by traditional banks. Online banking also allows its users to monitor their finances regularly. It al-

lows one to access his account history and past transactions from any location, as well as enables users to be in the know about any unauthorized transactions, so that they can be disputed in a timely manner.

2. Related Work

Recently, Deloitte, a top global provider of financial services, released an article highlighting the impact the relationship between a customer and a bank has on their business [3]. There are four key areas in which a bank must focus on; customer acquisition, service, relationship development, and customer retention. Within each of these areas there are a number of focal points in which a bank would require data in order to satisfy that requirement. You can see this visual below This is where the data modeling, similar to our study, comes into play.



This graphic was provided by Deloitte to illustrate the importance of maintaining each of the four key areas in order to ensure the longevity of a customer's satisfaction

The first step in enhancing the banking experience as well as the bank's success and growth, is to acquire more account holders. The areas Deloitte highlighted align with our motive for this study. Deloitte recommends that in order for banks to properly target and obtain a customer, the bank must understand the history of this customer. With the abundance of data a bank has access to, it is key to follow the correct insights and focus on what really promotes success rather than focusing on the wrong thing and wasting resources. This publication truly highlights the importance of getting to know each customer on a deep level in order to give the best experience possible. This relates to our work because it goes to show the significance of analyzing a customer in order to set your bank apart from the rest.

Deloitte goes on to stress the importance of building good relations with the customers, and bolstering customer retention rate. What is key in this publication is the utilization of data analytics. One can easily go through the process of programming and cleaning each customer's specific data but it is what you do with these metrics that is key.

While further analyzing studies currently being conducted in regards to online banking and customer data, it is important to note the programs these companies and banks are using. In September of 2019 an article by Jurgen Jürgenson, was published highlighting the rise of python as the one of the most 'in-demand' programming languages in banking [6]. Currently, leading banks like J.P. Morgan, Citigroup, and Bank of America use Python to analyze their customer's data. Our study and implementation of Python programming can be utilized to not only show the importance of focusing on your customer's background but also to convince banks that implementing our models is the most optimal way to do so. This article emphasizes the simplicity of Python in comparison to other programming languages. A compiled list of some of the best utilized Python libraries consisted of almost every library used in our study. Jürgenson especially focused on `NumPy` as one of the most fundamental supporting library. Similar to the publication by Deloitte in what data to use when analyzing your customer as to not make things increasingly difficult, how you choose to work through the data is just as important. When choosing a programming method, banks can be faced with a number of choices, however through Jürgenson's work and our own it is clear Python is the optimal language.

3. Proposed Method

In order to create the best model to predict whether a customer will use online banking, as well as assessing feature importance, we had to create multiple models and see which one performed the best. We decided to do this using K nearest neighbors and decision trees. K nearest neighbors is trained by "memorizing the training data" and then it predicts unseen data's class by class based majority voting based on the k closest points. Decision trees function by picking a feature that when the parent node is split results in the largest information gain. The splitting process is stopped if the child nodes are pure or the information gain is less than or equal to zero. Once each child node is pure or the potential information gain of the best split is less than or equal to zero, we have the final tree.

In addition to these base models we decided to incorporate multiple ensemble methods, specifically a bagged k nearest neighbor model and a bagged decision tree, as well as a random forest model. Bagging is the process of bootstrap sampling the training set n times and training a model on each of the n bootstrap samples. We then can make a prediction using all of these models. The final prediction is determined by majority voting between all the results of our n models.

The random forest model is an ensemble of decision trees. It is basically bagging with decision trees, except normally we would base the splitting criterion of the trees in the entire feature set. In random forests, we base the splitting

criterion of the trees on random subsets of the entire feature set. We specifically decided to use random forests so we could use it to asses feature importance. We also thought to use it as a baseline because random forests tend to work well with minimal hyper parameter tuning.

Since we are not using many algorithms, we decided to focus on tuning our hyper parameters. We used both random search cross validation and grid search cross validation. We focused on grid search cross validation due to the rigor of the test. We were not working with a massive data set so the computational expense of grid search cross validation was not a huge drawback. Cross validation is the process of selecting k "folds" within the data. We then create k models such that the data is trained on k-1 "folds", the kth fold is then used as a hold out/test set which is used to evaluate the model.

In grid search cross validation we exhaustively test each possible combination of provided hyper-parameters within our parameter grid. In randomized search cross validation, we randomly select hyper-parameter settings within our provided parameter grid. We found grid search to work the best as our data set was not particularly large. The computationally expensive nature of grid search was not a huge issue due to the low size of the data set.

Finally, we decided to use the random forest to assess feature importance. We felt this was critical to our overall goal of finding the features that are most important to predicting online banking. We also planned on plotting this to provide an easy visualization of feature importance.

4. Experiments

We performed the experiments described in the proposed methods section. Specifically, we tested the variability of the features and removed features with very high or very low variability, as well as features that would not generalize well. These included ID due to high variability, personal loan due to low variability, and zip code due to inability to generalize well. We then did a stratified train test split into training, validation and holdout sets. Then we elected to use a standard scalar for feature scaling. We tested all our models with both un-scaled and scaled features. We found that the performance of our K Nearest Neighbor and random forest models were boosted with scaled data, while the decision trees and other ensembles were unaffected. Therefore, we chose to keep the feature scaling. After cleaning and splitting the data we trained and evaluated our proposed models. The test set accuracy for each model is provided in the above table. One thing of note was that the base random forest was very over-fit. We then used a grid search selection on the random forest to increase its robustness and slightly boost the test set accuracy.

Method	Accuracy
Gridsearch Decision Tree	73.61%
Gridsearch KNN	68.75%
Bagged Decision Tree	73.61%
Bagged KNN	67.36%
Random Search Cross Validated Decision Tree	70.83%
Decision Tree	72.916%
KNN	72.22%
Random Forest	72.22%
Gridsearch Random Forest	72.97%

Table 1. Here is each method implemented in our study along with the accuracy.

4.1. Data set

We have used the data set provided on Kaggle called Bank Loan Modeling[5]. This data set is of 5000 observations and has fourteen variables. There are five binary variables which are Personal Loan, Securities Account, CD Account, Credit Card, and the target variable which is Online. There are seven numerical variables which are ID, Age, Experience, Income, ZIP Code, CCAvg, and Mortgage. Lastly, there are two ordinal variables being Family and Education. Of these thirteen features, not including the target variable, we decided to remove three of them. These three variables are ID, ZIP Code, and Personal Loan. ID was removed due to high variability, Personal Loan was removed due to low variability, and ZIP Code was removed due to lack of ability to generalize.

```
ID 480 480
Age 40 480
Experience 42 480
Income 102 480
ZIP Code 238 480
Family 4 480
CCAvg 95 480
Education 3 480
Mortgage 141 480
Personal Loan 1 480
Securities Account 2 480
CD Account 2 480
Online 2 480
CreditCard 2 480
```

List of all features used

4.2. Software

The computational software utilized in this project are Python and Jupyter Notebook. The `mlxtend` package, as well as the `Sci-py`, `NumPy`, `Pandas`, and `Scikit-Learn` packages will be used.

4.3. Hardware

The computer hardware which will support this project will be each member's individual laptop. All of the individual laptops ran on Mac OS. We used a Macbook Air with stock hardware as well as two Macbook Pros with stock hardware.

5. Results and Discussion

Going into this study we faced two focal questions. The first being, what is the best performing model on our data and the second being which aspect of the customer is the most important feature for a bank to appeal to.

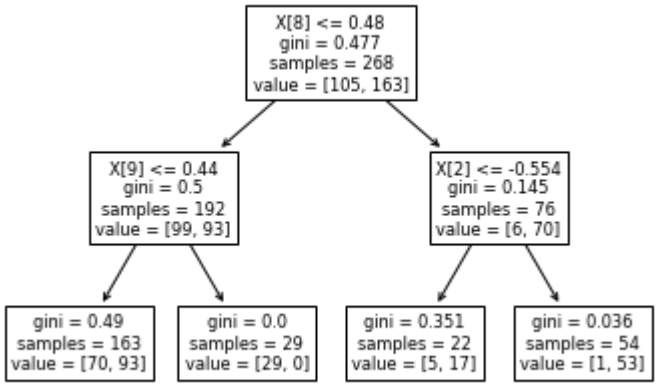
We began using a base K Nearest Neighbor, Decision Tree and Random Forest model. We used our random forest as a baseline of sorts due to the high performance of the model without much parameter tuning. With our non-crossed validated models, we noticed that the accuracy across the Decision Tree and K Nearest Neighbor were quite high, with the base Decision Tree being slightly more accurate on the test data. However, we noticed that our Random Forest had a relatively high test accuracy, yet was extremely over fit. The high test accuracy was to be expected as random forest tends to have high performance with minimal parameter tuning.

We then decided to do a randomized cross-validation search on the decision tree. We decided to begin with randomized search cross validation because it is not as computationally expensive as a more exhaustive grid search cross validation. This way we could test a wide range of hyper parameters. Surprisingly, our randomized search cross validated Decision Tree ended up with a lower test validation than our base decision tree. The better tree can seen below as Graph 1.

Graph 1. Random Search Decision Tree

After using the randomized search cross validation on our decision tree, we then decided to use the information gained about hyper parameter values on a more exhaustive grid search cross validation. Our grid search cross-validated Decision Tree yielded the highest Test Accuracy, while surprisingly, our grid search cross-validated K Nearest Neighbor model had a lower accuracy than the base K Nearest Neighbors. We are unsure why this happened, but we assume it is related to the parameter grid provided to the cross validation object. We then used grid search cross validation on our Random Forests model. This resulted in a slightly higher test accuracy, but more importantly, it decreased the over-fitting of the model significantly. We went from a test set accuracy of 100% to 69%.

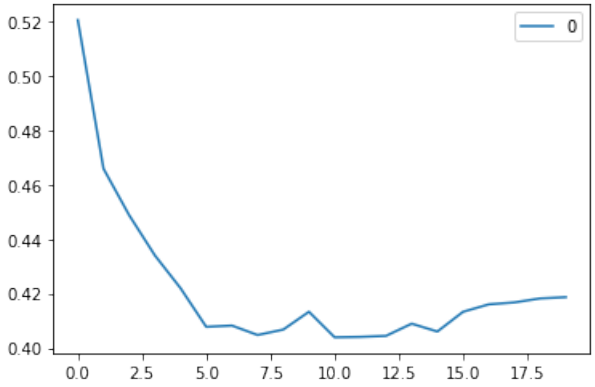
After cross validating with grid search and random search, we decided to create more ensemble methods. We settled on a bagged K Nearest Neighbors and Decision Tree as we were the most comfortable, with those. Our bagged K Nearest Neighbors performed slightly better than our grid searched one, and our ensembled Decision Tree matched the test accuracy of our grid search cross validated one. We ended up with our grid searched Decision Tree and our bagged Decision Tree having the highest test accuracy of 73.61%. Their training accuracy's were marginally different, thus we decided to call both of them the best models.



Graph 2. Grid Search Cross Validation Decision Tree Initially, we thought cross validation and ensembling would lead us to the best performing model. Our reasoning behind this hypothesis was that when utilizing ensemble methods you are increasing the likelihood of a better prediction and performance. This is because, by ensembling, we combine the power of multiple models trained on different subsets of the data and features. The predictions made by these models are then amalgamated into a single prediction, accounting for all the models within the ensemble. By ensembling, we reduce the variance without increasing the bias.

We expected cross validation to perform well too because it allows for exhaustive testing of hyper parameter combinations. From our computing we found our best performing models from Gridsearch Decision Trees and Bagged Decision Trees yielding a

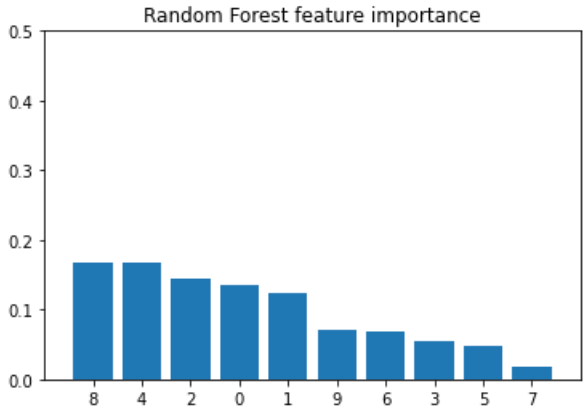
73.61% accuracy and 73.61% accuracy, respectfully.



Graph 3. RMSE of K values in KNN, K values on X axis, RMSE on Y axis

The results we found were to be expected with 2 exceptions. Both of these exceptions relate to the K Nearest Neighbor algorithm. Both our bagged and grid searched K Nearest Neighbor performed worse than our base K Nearest Neighbor with k=6. We are unsure why this happened as we gave the grid search cross validation object quite a wide range of hyper parameters to exhaustively test. We thought we may face issues with over fitting, as seen in our Random Forest, this left us surprised as this is not a common issue with Random Forest especially with such a high number of trees. Thankfully, we were able to resolve this issue by cross validating the random forest with a grid search.

Variance is reduced and accuracy is improved, without increasing bias, during bagging which led to a good model. Additionally we found the Gridsearch Decision Tree to yield an equally high accuracy. This also came as no surprise because the hyper-parameter tuning allows for each combination within the grid to be evaluated.



Graph 4. Feature Importance Graph. Legend for x-axis: (0: 'Age', 1: 'Experience', 2: 'Income', 3: 'Family', 4: 'CCAvg', 5: 'Education', 6: 'Mortgage', 7: 'Securities Account', 8: 'CD Account', 9: 'CreditCard')

Our second goal was to find the most important features of a customer for the bank to learn about in order to enhance business performance. The Graph 3 pictured above plots our feature importance. We found the top three features a bank should be focusing on are; 'CD Account', 'CCAvg', and 'Income'. The CD Account being the most important feature refers to whether or not the customer has a certificate of deposit account with the bank. CCAvg, the second most important feature, measures the average spending on credit cards per month. The third most important feature is Income which measures the annual income of a customer. These results are not overwhelmingly shocking. A certificate of deposit is a way for a customer to obtain an interest rate premium as long as they promise to leave a certain amount of money deposited in the bank over a period of time. This measures a customers loyalty to the bank in turn making it not all too surprising that this should be the number one feature for a bank to prioritize when hoping to enhance a customers experience. The average amount spent on a credit card each month also isn't too surprising as it shows how valuable a credit card is to a customer. If the customer is barely spending money on their credit card they likely feel less attached to their bank rather than those who spend a lot and require high return rates and benefits. Finally, the third feature importance is income which the reasoning for this likely aligns with that of average credit card spending per month. The more money there is, the higher the stakes.

The way in which we came to these results regarding feature importance is through Random Forest. At first we implemented feature importance on our Random Forest model. We saw that the Random Forest was very over fit which led us to have to rethink our approach. Once we cross validated the Random Forest and then utilized feature importance we found different results than the first time. It is really important that we cross validated because if we hadn't, we would have come to an incorrect conclusion. Once we had the properly cross-validated Random Forest feature importance we plotted the findings, as seen in Graph 3, to help better visualize the results as well as make the delivery of the results easier to others.

6. Conclusions

The analysis presented in the report aims to find the best model as well as feature importance. We sought to fit a machine learning algorithm to accurately assess online banking as well as the features of the biggest importance in this assessment. We concluded that the GridSearch Decision Trees and Bagged Decision Trees perform the best of all the models with an overall test accuracy of 73.61% for both algorithms.

In terms of the most important features in predicting whether a customer would elect for online banking, we have concluded that there are four features that are of most importance. These four features, listed in ordered of impor-

tance are CD Account, CCAvg, Income, and Age. CD stands for certificate of deposit account, and CCavg stands for the average amount spent on a credit card monthly. These four features can help banks identify customers who may prefer to use online banking. They can then be targeted by advertising to encourage adoption of online banking. By introducing customers to online banking, the bank can increase customer satisfaction, and bring in a new, younger crop of account holders.

6.1. Future directions

In terms of what should be done going forward to improve the experiment, there are a couple of aspects that could be brushed up. First, we would suggest that more time is spent cleaning up the data. We did in fact spend a great deal of time data cleaning, however, we feel that with more time and focus on data cleansing, the experiment could yield more accurate results. Secondly, our results suggest that Decision Tress as well as Ensemble Methods produced the best results, thus going forward we would suggest spending less time on K Nearest Neighbors. Additionally, moving forward we would suggest focusing more on trying to raise the test accuracy.

7. Acknowledgements

For this project, we used the dataset, Bank Loan Modeling, published to Kaggle by Sunil Jacob [5]. We additionally appreciate all the guidance received from Dr. Sebastian Raschka.

8. Contributions

All group members worked collaboratively throughout the course of the project, from the initial data review and clean up to drawing conclusions from the results. Each group member contributed in all areas of this investigation, contributing an equal amount. When finding the data the three of us worked together to establish a topic and find a proper data set to work with. During the computational step, each member had a different task. Jessica Ostroff created Decision Tree Plot 1, Decision Tree Plot 2, the base Decision Tree, and the Base K Nearest Neighbors Model. Claire Westenberger worked on: feature scaling, a looped K Nearest Neighbors, testing k values of 1-20 with a plot of the RMSE values, both our base and grid searched Random Forests as well as creating a graph for feature importance on our base Random Forest. Manav Kalathil worked on: testing feature variability to find candidates for removal, the removal of said candidates, splitting the data, a grid search cross validated Decision Tree and K Nearest Neighbors, a bagged Decision Tree and K Nearest Neighbors, a randomized search cross validated Decision Tree, and the feature importance graph for the grid search cross validated random forest. Finally for the project report we split up the report by section and worked with each other to contribute knowledge and edit each other's work.

References

[1] 18 online banking statistics you need to know in 2021. September 2021.

[2] Online banking users worldwide in 2020 with forecasts to 2024, by region. October 2021.

[3] Deliotte. Finally: Customer analytics for banks. 2011.

[4] B. Gran and D. Foreman. Why banks need to do more to help customers become financially successful. *Forbes Advisor*, March 2021.

[5] S. Jacob. Bank loan modelling. August 2018.

[6] J. Jurgenson. Python is crushing banking - choosing the right tech stack for your fintech solution. September 2019.

[7] S. Singh and R. Srivastava. Understanding the intention to use mobile banking by existing online banking customers: an empirical study. *Journal of Financial Services Marketing*, 25(3):86–96, 2020.