Duc Tran Greenwell

Hasib Azami

Johan Sjoden

Elizabeth Romero

Kevin Eddy

Professor Sabbaghi

Capstone; Autumn 2021

December 11, 2021

**Customer Churn Prediction**

# Abstract

As with any business, overall success is directly affected by its ability to acquire and retain customers. Achieving this need has become increasingly more difficult in recent years due to rising competition, and market structuration. In the U.S. alone, the banking industry spends nearly 1.5 billion dollars annually to acquire new customers. The bulk of these costs can easily be avoided by focusing on customer loyalty and improving the longevity of their current customers.

Aside from the rising costs banks are seeing with acquiring new customers, it is becoming more apparent that maintaining customers increases profitability. This is true because long-term customers tend to produce more profits as they tend to have multiple accounts, Verbeke et al (2011). “According to Nie et al (2011), a bank can increase its profits by up to 85 percent by simply improving their retention rate by a mere 5 percent”, Keramati (2016). This contributes to banks' rising attention to churn management and their efforts to build such prediction models.

With so much to gain, more businesses across industries are drawing their attention to churn management. Provident Credit Union is no different. In this paper, our team uses data provided by Provident Credit Union to analyze customer behavior and to build a variety of machine learning models to effectively predict customer churn. It is our hope that through our analysis we will be able to successfully identify key features that contribute to future churn and give Provident Credit Union the necessary tools to take preemptive actions to retain customers.

Objective: we seek to use key data features of churned customer characteristics to build a reliable model to predict which of Provident Credit Union’s customers will churn, why people churn, when they will churn, and an app to interpret the prediction outcomes to help the client gain an insight to reduce the churn rate, accordingly, to increase profitability.

# Introduction

With the banking industry becoming more saturated, banks are relying more heavily on customer retention strategies to increase profitability. On average, a bank can increase its profits up to 85% by retaining just 5% of its pre-existing customers (Prediction of Customer Churn in Banking Industry, Charandabi, 2020). Therefore, analyzing customer behavior data and monitoring their financial activities has become a critical component for banks to predict their customers’ future actions and help reduce profit loss.

To date, most of the work done in churn prediction has centered around sampling strategies, feature engineering, and transforming data into a more organized and meaningful structure. With the customer data provided by Providence Credit Union, we will develop and validate our models following the similar structure of previous churn prediction methods. In doing so, we aim to accurately capture customer behavior for churn prediction using the Logistic Regression, Survival Analysis, Decision Tree, and the Random Forest models.

The reasons behind choosing these models are because we found from various studies that the Logistic Regression model was the most popular one. Logistic regression is easier to implement, interpret, and very efficient to train[[1]](#footnote-1).

Survival analysis has been widely used in clinical research. This model enables researchers to estimate the survival probability of an observation or a group of observations at a given time t0 until the event occurs at tk (i=0...k). Beside providing survival estimates, the exponential coefficient outputs from Proportional Hazards Ratio can help explain the contribution of individual variables[[2]](#footnote-2).

Compared to other algorithms, Decision Tree can learn at a granular level, and is very intuitive and easy to explain. It also does not require normalization nor scaling[[3]](#footnote-3).

Finally, Random Forest algorithms produce high level of accuracy while being capable of handling large dataset efficiently[[4]](#footnote-4). From these collective algorithms, we believe that we will be able to reach a high level of model using the data provided.

All the work has done so far undeniably a long journey. However, results are noncommunicable without proper interpretation since the prediction outputs are very heavily technically interpreted, management members would find it difficult if not misinterpret the results unless they come from similar backgrounds, which is a barrior for any institution. Besides, it would take a good amount of time to translate the

outcomes into business languages. To help our client avoid the nuisance and cost, we built an app called UI (user interface) to print out the output from our model predictions.

# Literature Review

In the research paper “Developing A Prediction Model for Customer Churn From

Electronic Banking Services Using Data Mining”[[5]](#footnote-5), Abbas Kermati, Hajar Ghaneei, and Seyed Mohammad Mirmohammadi used data mining techniques for predicting customer churn within the banking industry. The aim of their research was to identify the appropriate features that represented churners from electronic banking services. While the bulk of their data was mined from open data sources, the length of customer association, and customer complaints were provided by participating banks. They used both backward and forward elimination methods to detect feature subsets, however found that backward elimination ultimately performed better. Through this process they found that the career variable was redundant and removed it before applying the decision tree model.

In the end their decision tree model returned 5 combinations of features that indicated future churners, all of which showed customers that had less than 0.5 mobile and 1.5 online transactions per month. They ended their research with the recommendation of further qualitative research be done around these features. Better understanding customer intentions will prove beneficial to banks who want to meet customer requirements to prevent them from churning.

Most of the work done in churn prediction has focused on sampling strategies, feature engineering and supervised modeling over a fixed period of time. In the research paper “A Dynamic Classification Approach to Churn Prediction in Banking Industry”[[6]](#footnote-6), Hoiyin Christina Leung and Wingyan Chung aim to develop and validate a classification approach to using customer behavior data to predict churn. While few studies in churn prediction have explored using dynamic predictors over longitudinal data, Leung and Chung are seeking a way to do just that. Their approach uses dynamic time-series predictors over multiple time periods with rare event detection to build a model that will enable accurate and long-term predictions.

Using a 3-year transaction dataset sourced from a local Florida Bank, Leung and Chung found that their model showed training data extracted over 6 months performed better than 4 months, however a 2-month prediction window has the benefit of avoiding any rapid decay of accuracy. Their study used multiple time periods to test if any marginal gain in accuracy would decline with additional data. Overall, Leung and

Chung found that their approach to predicting churn could be limited by the lack of independence around observations of the same customers. They recommend future research be done to address this issue and improve static predictors.

The profitability of long-term customers has been noted across industries and companies are becoming more proactive in establishing these long-term relationships. In addition to creating and improving loyalty programs, many companies have been looking to machine learning techniques to better detect customers with high propensity to churn. In the research paper, Machine Learning for Customer Churn Prediction in Retail Banking, Joana Dias, Pedro Godinho, and Pedro Torres present their work to predict customer churn in retail banking. Their methodology consists of using 6 machine learning models to predict up to 6 months of who will and when will churn. Each model will predict which customers will churn in 1 month, 2 months, and so on until reaching 6 months.

Using a dataset representing more than 130,000 customers of a retail bank across 2 years, Dias, Godinho, and Torres created different datasets to train their models using rolling time windows from one to six months7. To use as much data as possible, they decided to use leave-one-out cross validation to test the models and calculate their accuracy. As for their models, they used random forest, support vector machines, stochastic boosting, logistic regression, classification and regression trees, and multivariate adaptive regression splines. From these models, they found that stochastic boosting performed the best when predicting 1 - 2 months ahead. They also found that the most important feature when predicting customer churn was how many products each customer had during the past 3 months, along with any debit or credit cards at another bank. Overall, their study only highlights customers' motivations to churn at a high-level and doesn’t truly characterize each customers’ unique motivation to churn.

## Data

The data used in this project is from Provident Credit Union, including all customer interactions with the bank from 1951 to October 31st, 2021, with more than 16 million records of more than 151 thousand members holding multiple account types. Customers’ identification was anonymized to maintain confidentiality.

7 *Machine Learning for Customer Churn Prediction in Retail Banking*

Joana ias, Pedr Godinho, and Pedro Torres

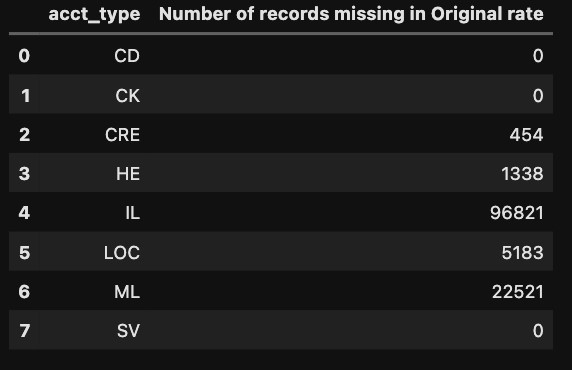
Initial data include 6 transactional files (members, Balance history, Deposit, Visa, Loan Transaction, Moving by Zip codes), 5 survey files (Second email survey, New Management Survey, New Member Survey, Branch Survey, New Consumer Survey), 1 promotion file, and 1 member’s fee file. Deposit transaction file includes deposit account type: SV (savings), CK (checking), CD (certificate). Loan Transaction file includes loan account type: CRE (commercial real estate), HE (home equity loan), IL (installment loan), LOC (line of credit), ML (mortgage loan). Visa file includes both credit and home equity loan if the amount borrowed greater than $35,000.

### Preliminary data examination

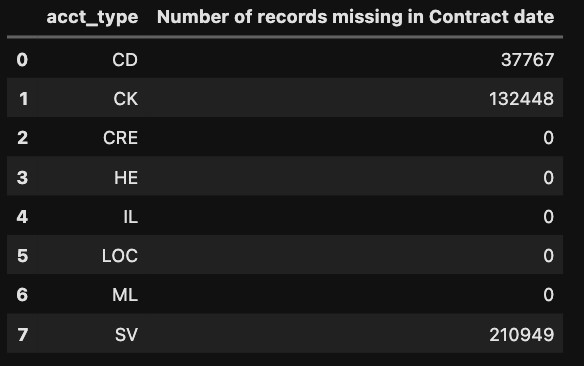
Total original records by account types



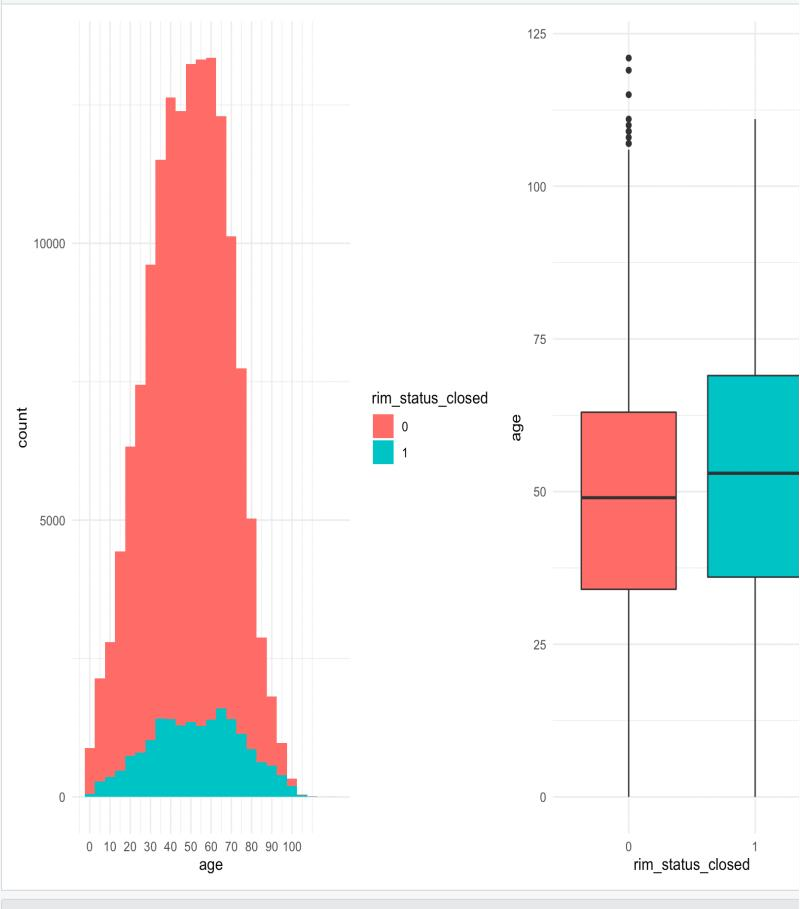
Null value checking in Original rate column



Null value checking in contract date column

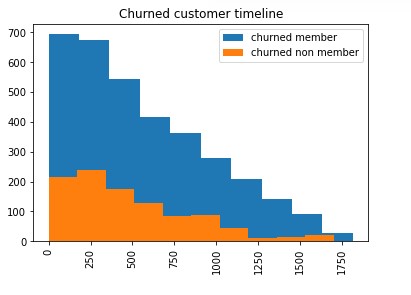


Original data distribution by ages

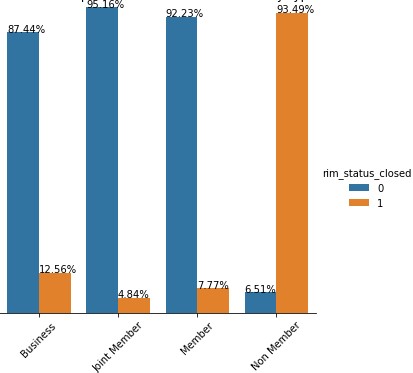


To overcome the challenge that it was difficult for regular computer to store all the files at once, these files first went through a cleaning process individually, initial evaluation with statistical analysis, transformation, and visualization, then the cleaned files were merged into a master file for grand data transformation and variable selection steps. For variable selection, Chi square test was applied. The final set of attributes consisted of age, account types (CD, IL, ML, VISA), mrm, term, original rate, account duration, branch rating, member rating, & contract rating, moving time), enrollment/effective dates along with closed/maturity date, account activity, and current and original rate, and current and original balance.

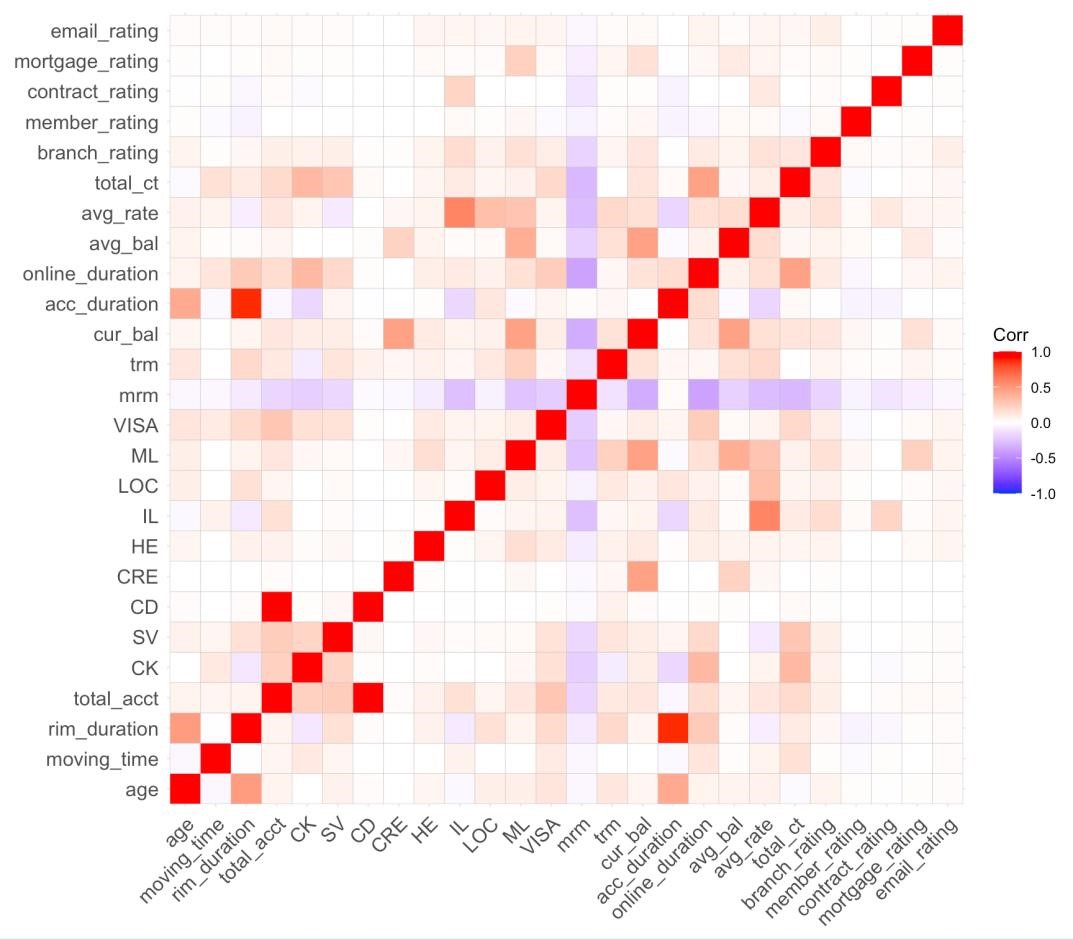
Distribution of churn vs non churn by age



Churn vs. Non-churn by account type



Confusion Matrix amongst variables



## Churn Definition

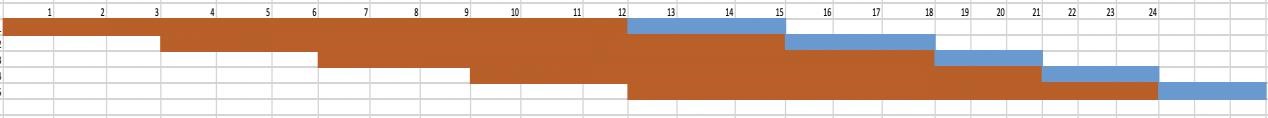
According to Provident Credit Union, the definition of churn is when a membership cancellation happens. However, for data integrity and model performance reliability, we only selected members’ records whose status were classified as active and closed and excluded the rest.

The challenge we faced with predicting customer churn was that customers with similar characteristics may not all churn at the same time and their motivations to churn may be unique to them. Most importantly, the dataset was highly imbalanced with most customers representing non-churners.

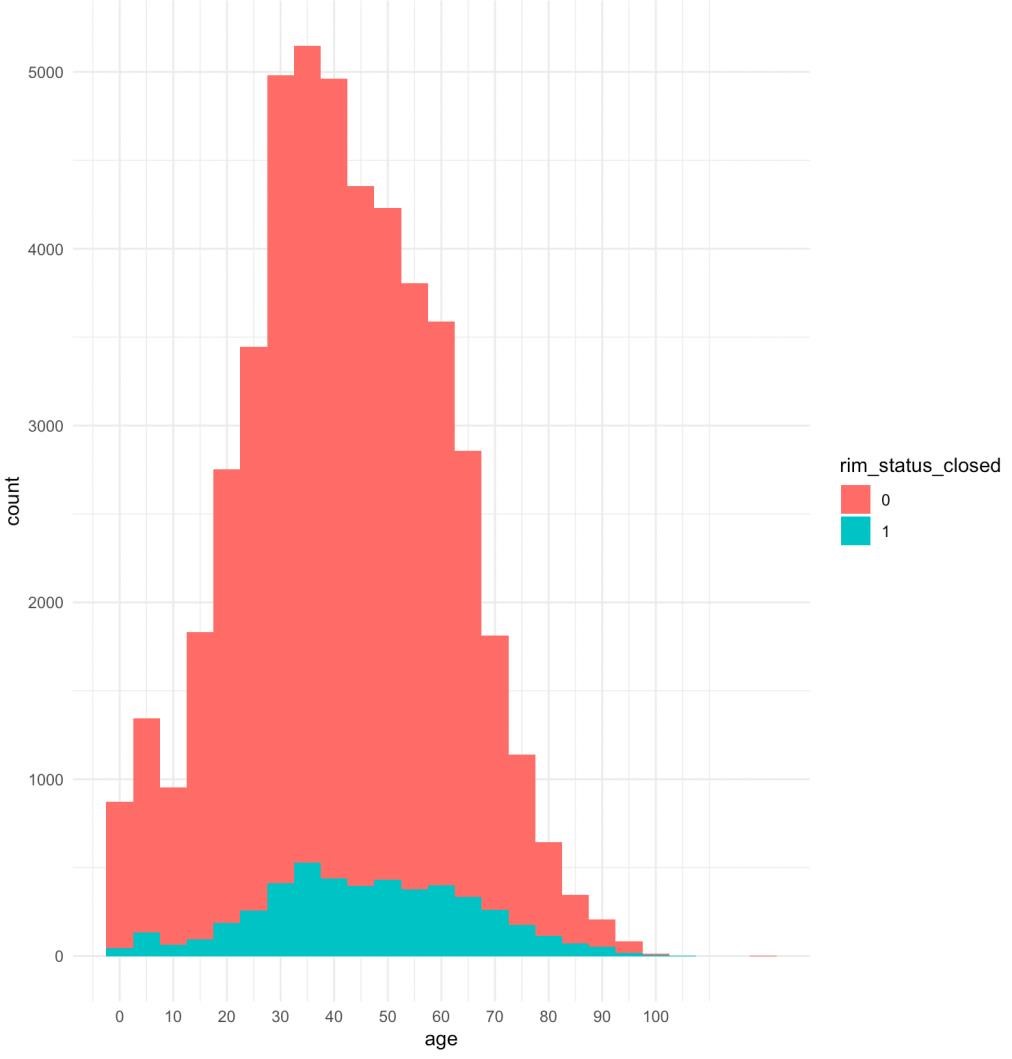
## Data Structure

After the data was transformed from long format into wide format, we selected 5 most recent year data and constructed into stepwise structure using rolling window method. The records were rolled each 12 months in an observation window which constitutes 4 periods, then after each observation window followed by a 3-month performance period. During the observation periods, both closed accounts and closed member status were excluded. In the performance period, however, we included this piece of information for the model to predict who would churn within the next 3 months. By constructing data this way, a member will contribute multiple records to the global data set until the churn occurs.

### Rolling Window Structure



Distribution with of churn Vs. non churn by age (rolling window data)



Once the desirable dataset was achieved, we trimmed outliers, and followed was a Chi-square Test. Out of 28 variables, we dropped 14 variables that are not statistically significant. These variables are CRE, Home Equity, Line of Credit, current balance, original amount, online duration, missing payment, average balance, average rate, total amount, total count, available balance, mortgage rating, email rating.

Variable Selection

Calendar

Description automatically generated

With a clean dataset in hand, we stratified the data set with a ratio of 70% of train data and 30% of test data. Yet, bias was still present since the minority class was extremely small relative to the majority class while the algorithms were accuracy driven i.e. they aimed to minimize the overall error to which the minority class contributed very little because the algorithms assumed that the data set had balanced class distributions. They also assume that errors obtained from different classes have the same cost.

Therefore, the very last step before fitting data into our models, we conducted a test using Logistictics Regression for over sampling, under sampling, both, and ROSE techniques.

With an oversampling method, it replicated the observations from minority classes to balance the data. An advantage of using this method is that it leads to no information loss. The disadvantage of using this method is that, since oversampling simply adds replicated observations in the original data set, it ends up adding multiple observations of several types, thus leading to overfitting.

In contrast, the number of observations from the majority class were reduced to make the data set balanced in under sampling technique. Apparently, removing observations may cause the training data to lose important information pertaining to the majority class. However, this method was appropriate for a huge data set, plus it reduced the number of training samples and helped to improve run time and storage troubles.

The both sampling method is the combination of the over and under sampling method.

Lastly, ROSE (Random Over-Sampling Examples) aids the task of binary classification in the presence of rare classes. It produces a synthetic, possibly balanced, sample of data simulated according to a smoothed-bootstrap approach.[[7]](#footnote-7)

Listed above techniques produced rather close overall scores. While the undersampling technique has the same F1 score and AUC as oversampling and lower AUC score compared to the both technique, we decided to adopt the under sampling dataset for training our models given the fact that we had a huge data set.

Subsampling Method & Scores



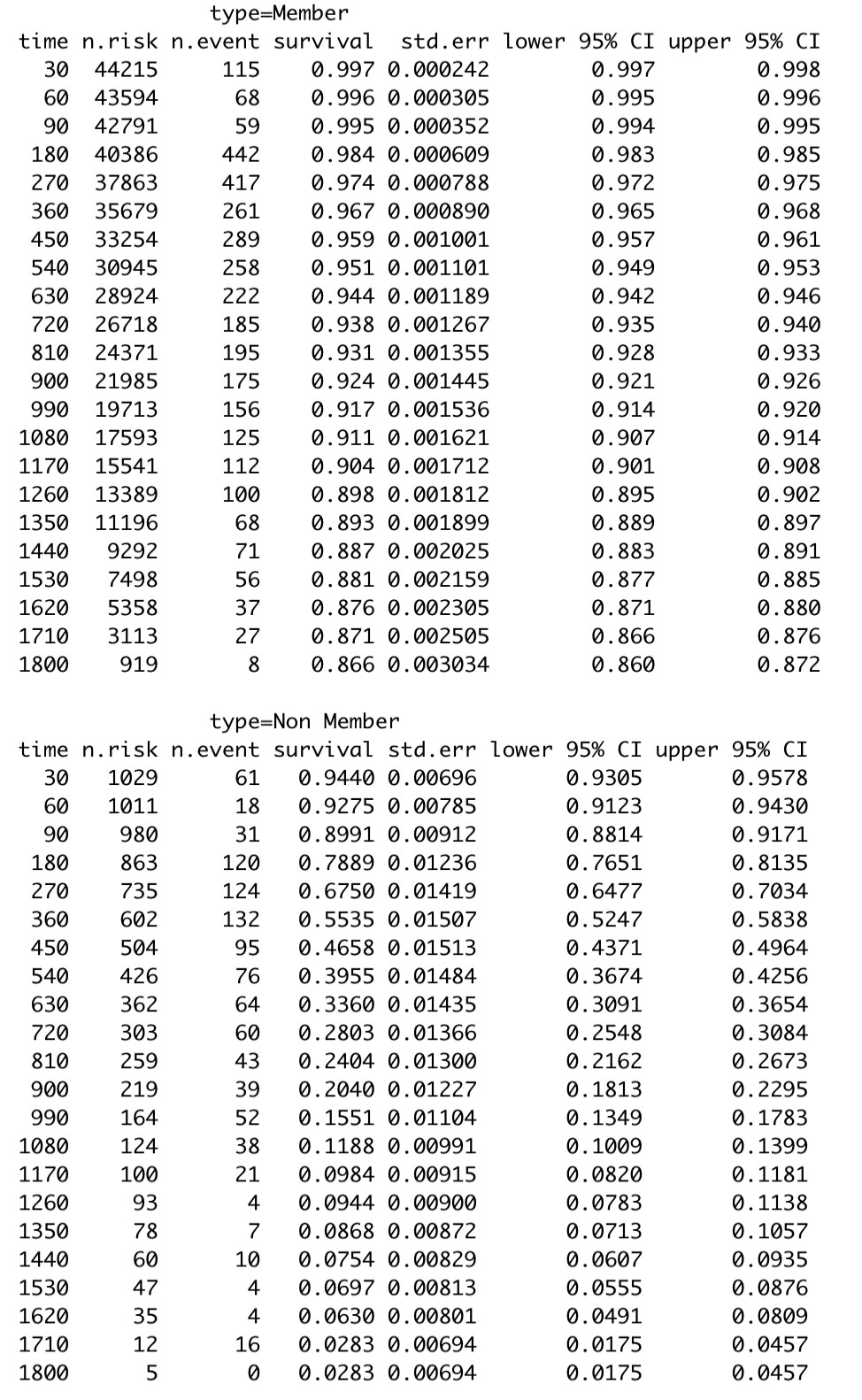
## Models

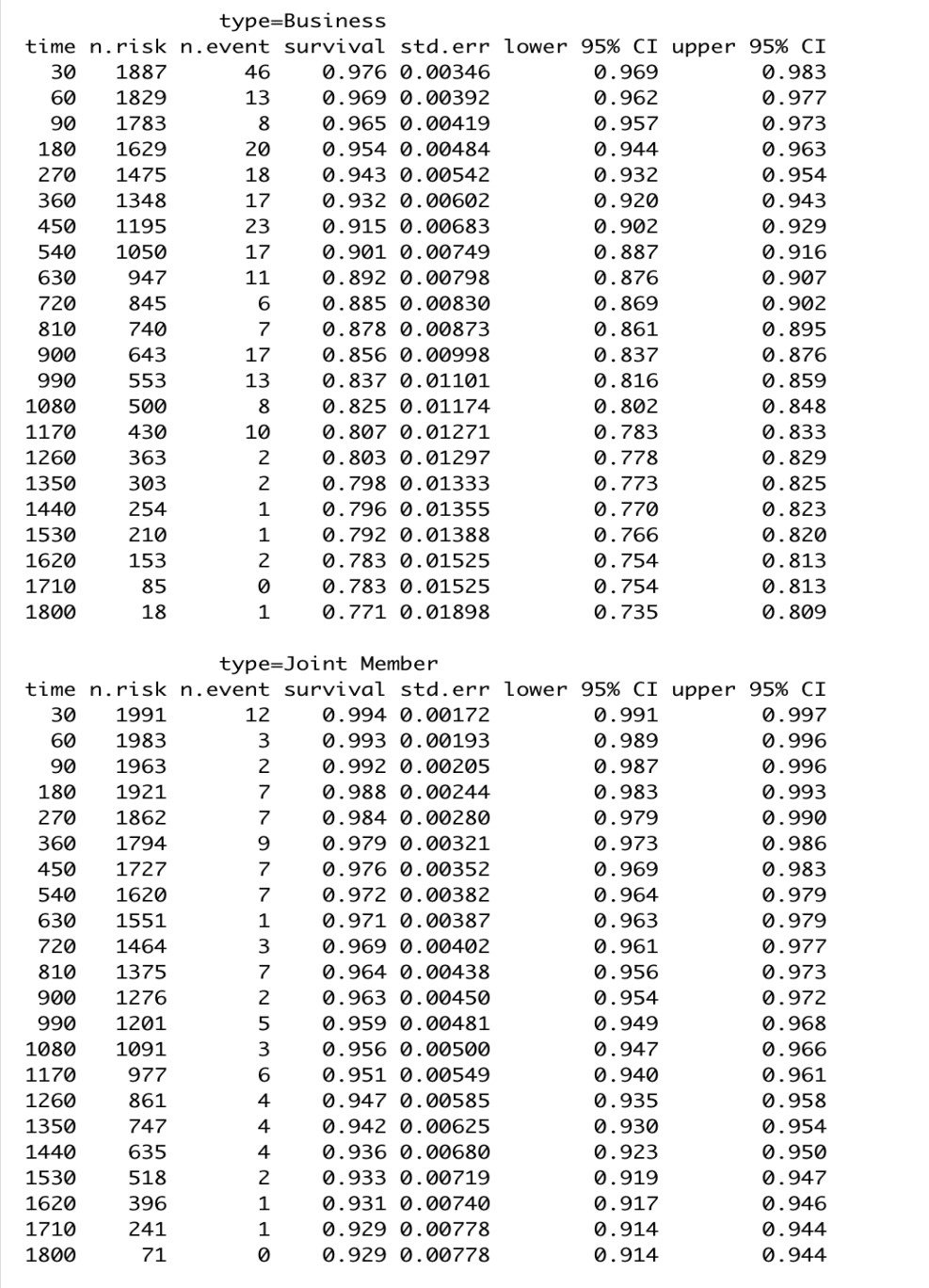
### 1. Logistic Regression

Predicting customer churn is at its core a classification problem, and this is the most widely used. Therefore, we decided to use a Logistic Regression model. This model returns a quite high score. Putting the accuracy measure aside, its F1 score is .028 and AUC is 0.894. Thus, this is a high performer.

### 2. Survival Model

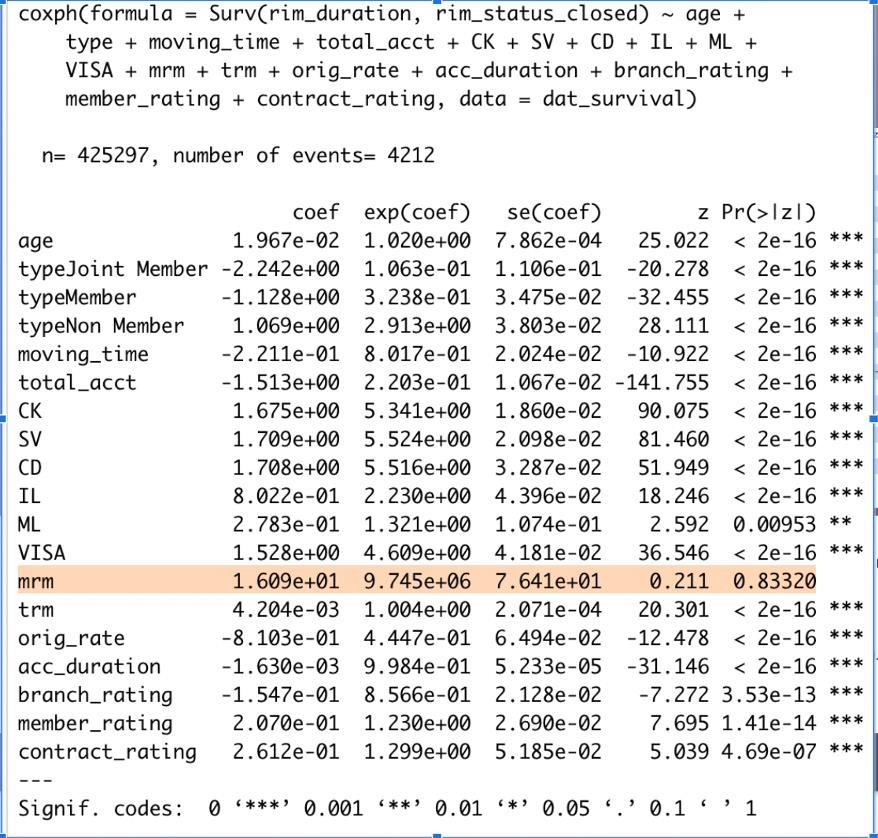
Next, we chose to use the Cox Proportional Hazards Model because it is one of the most commonly used in survival analysis. This model is used to identify the association between an outcome and one or more predictor variables. For example, we are looking for the association between churners and potential categorical predictor variables like “Type” or “Available Balance,” or any other significant variable(s).



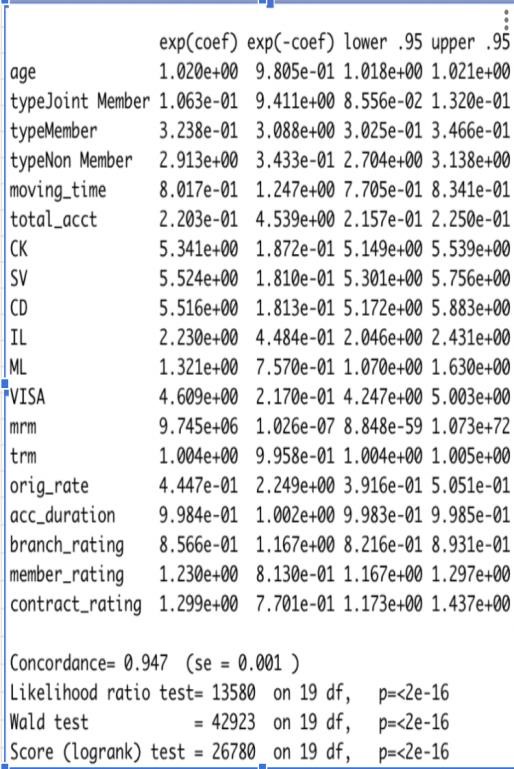
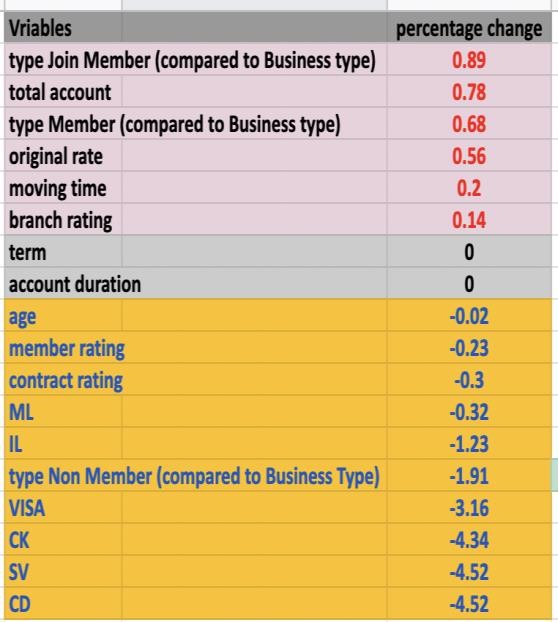


The Cox Proportional Hazards model indicated that all of the variables but the member relationship scores were statistically significant at 99% confidence level. We also got a concordance index of 95%. With this index and after a rigorous data transformation process, we believe this model is a good one for predicting. It’s worth noting that since we were faced with time constraints, we did not take additional steps to verify if this model violated the assumption of linearity or not. Therefore, further investigations needed as a safeguard before mass application of this model.

Cox proportional hazard model



Interpretation of hazards ratios Hazards ratios



According to the Proportional Hazards ratio, we can conveniently understand insights of contributive factors as well as the causes for membership churn.

More specifically, a member who doesn’t have a CK and savings account will be more than 4.5 times more likely to churn than someone does, holding all other factors constant. Someone who doesn't have a checking account will be more than 4.3 times more likely to churn than someone who owns a checking account, holding all other factors constant. Same pattern, someone who doesn’t open a credit card account will be more than 3 times more likely to churn than someone who owns a credit card account, holding all other factors constant. A non-member type will be 1.9 times more likely to churn compared to a member under business type, holding all other factors constant. And either being older or rating lower scores in member rating or contract rating will negatively impact the churn decisions, holding all other factors constant.

On the opposite end, while account term and lifetime of an account don’t impact the churn decisions, total number of accounts a customer have, member type, original rate, moving or not, and branch service ratings correlatively impact a customer’s churn decision. This piece of insightful information is a powerful tool to our client for strategic planning.

### 3. Decision Tree

We picked Decision Tree because it is a binary predictive model in which the data either satisfies or doesn’t satisfy the conditions established. The decision tree can learn about a training set to the point of high granularity. This model produced very high sensitivity and recall; however, the rest metrics were all lower than Logistic and Random Forest.

Although this model performed well it isn’t reliable because it causes overfitting in the learning process as the training dataset becomes overly optimized which makes the test set by default have error.

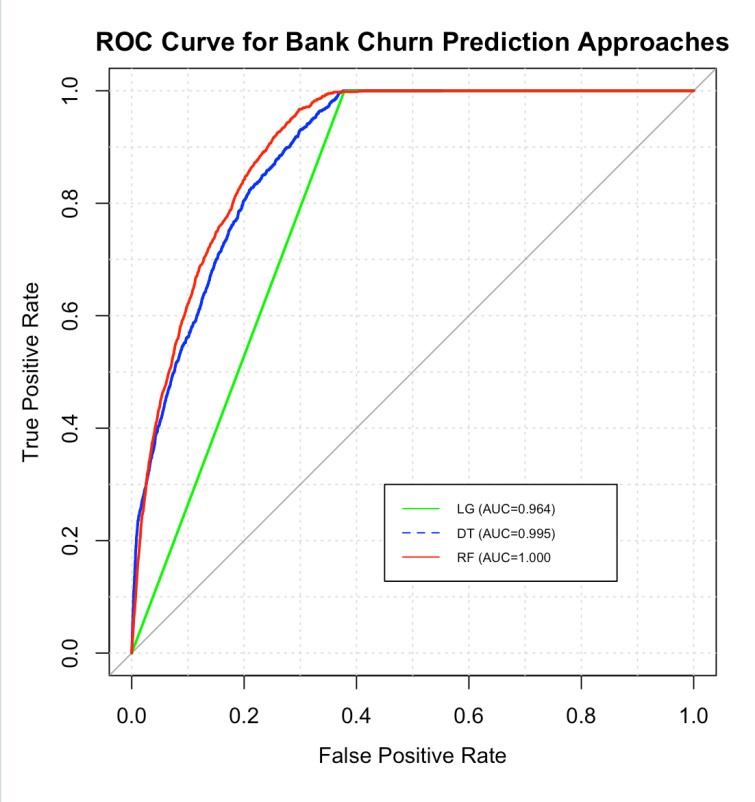
### 4. Random Forest

This model has been known for its highly accurate and reliability. The reason we choose the Random Forest model is because it can produce a high accuracy level. The result from this model confirmed the fact that it was the best candidate for application with the highest AUC score at 0.905.

Logistic, Decision Tree, and Random Forest score chart



ROC curve for the three models



## Conclusion

Taking all above analysis into consideration, we decided to adopt the Random Forest model for the client’s business application.

### Recommendation

Overall, churn analysis is a powerful tool and continues to evolve across industries.

While our analysis only scratches the surface of truly understanding Provident Credit Union’s customers, there is still a significant amount of work that can be achieved from our findings.

**Areas of Focus**:

We recommend that Provident Credit Union focuses more on the customer journey and better understand member and non-member needs and interests. Our main recommendations are broken into areas as follow:

* CD, SV, and CK

We found that members and non-members are significantly less likely to churn if they have a CD, checking and savings account. Thus, one way to encourage customers to stay longer would be to create proper marketing campaigns for these areas.

* Loan Conversion

Loan is the next area to prioritize resources to invest in encouraging customers to stay longer. Such marketing campaigns that specifically target those without loans are proper.

* Member type
  1. Member types are also the areas of attention in the effort of reducing churn rate. There should be more focus in attracting more business type members.

* Customer Service
  1. Now we know why customers churned, additional effort Provident Credit Union should take is to increase training the front end staff to better serve customers.

* Incentives
  1. It is without a doubt that the banking industry is highly competitive. The rising costs associated with acquiring new customers makes retaining current customers increasingly more important. Incentives will encourage customers to take action on products and services that they may have been considering for a while. Pairing personalized incentives at key points throughout a customer’s life cycle will ultimately increase loyalty and engagement while steadily positioning Provident Credit Union as their primary banking institution.

## Application

So, what does this mean, how can Provident Credit Union implement churn analysis into their business?

After training the models, evaluating them, and selecting the Random Forest for business application, we saved this model. Next, we installed a UI app which includes a front end (UI) and a server, and connected the trained model into this app. We also prepared new data with the same format as the previous used for the train set. The prediction results then would be displayed on the front end of the UI after our Random Forest model spitted out the results.

## Client Feedback: Unknown Individual Contribution

**Hasib Azami:** This was one of the most challenging projects I have worked on while I was in school but the learning experience was amazing especially working with such a talented cohort and team. If I had the chance to do this again, I most definitely would if it means gaining the knowledge and experience working as a Data Scientist at Provident Credit Union. Early on the project in the data cleaning process, I helped create dummy variables for the data columns in order to get a binary output for our machine learning models since we can’t use columns with strings. I also helped develop models in the early stages using the cleaned dataset in order to test it out using a Logistic Regression Model and check its recall and precision so we can utilize that data through R for our statistical analysis. I also helped develop the survival analysis models such as Cox Proportional Hazards Model and Kaplan-Meier Analysis Model in order to check the time at which members will churn using the right censored data column that I created using rim\_closed\_dt. This was a tedious method that I worked on but found many helpful resources online that assisted me with developing the survival models as well as the visuals and plots.

**Duc Greenwell:** This project gave me an opportunity to learn how to work on a real business scenario where data given was not as clean as homework assignment. I worked on every part of this project.

There were different types of challenges I want to mention here.

1. Learning curve

Dealt with huge raw data sets which covered a whole business background that needed a good amount of time to comprehend.

1. Learning along sides with working on the project

Survival analysis and cox hazards ratio were new, and I had to do extensive research to be able to understand, interpret and explain to my teammates. I had to self-teach most of the R parts to be able to complete the project.

1. Time constraints

Initially I used Python to connect the data, but the data came out incorrect. Considering the need to learn in either languages, Python or R, while the requirement was to implement models in R, I turned towards R with limit prior knowledge of it. It became stressful given the short amount of time for this project.

**Johan Sjoden:** My Part to this project was doing all data exploration, creating charts and descriptive statistics for presentation and the report as well. Also, it was to provide feedback on how the models were developed and correct any actions that were incorrect. Furthermore, making sure we had the correct variables for our models and making sure it was ready to feed to the models we had. This also included visualizing the variables that were significant (chi\_quare test) and importance. This was in a nutshell doing the feature engineering for our project/presentation. I believe we really came together as a group at the very end and were able to work together as a cohesive group to come up with a solution.

**Elizabeth Romero:** To start this was one of the most challenging projects throughout the program. It tested our overall knowledge and comprehension of everything we learned throughout the program and how to put it into practice. I contributed in various ways to this project including setting up the team meetings, some data cleaning to understand the data, creating a logistic regression model, helping write some parts of the report as well as reviewing it, and providing support wherever it was needed. I learned a lot during the last eight weeks on the real-world applications of data cleaning, feature engineering, data exploration, and model building. I also learned the importance of project management and how it can impact team organization, communication, and the project. Overall, it was a very good learning experience for me.

Thanks to the Provident Credit Union team for all of your support.

**Kevin Eddy:** This project most definitely did not come without challenges, but through hard work and perseverance we were all able to pull through and provide some valuable insights. My contribution for this project was to help bring it all together and outline the key takeaways. Early in the project I created a Gantt chart to breakdown each task along with their expected completion dates. I also created a story in Tableau consisting of several dashboards that shared visuals of the different accounts and account types of members and non-members had at Provident Credit Union. After all of our models were completed, I used what we found to establish recommendations where Provident Credit Union could use to reduce customer churn and increase retention. In addition to all of this I created the slide deck and wrote the final report, with the exception of a few parts. I also contributed to our initial research of customer churn in the banking industry and potential models to use.

## Cited Work

1. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logisticregression/
2. https://youtube.com/watch?v=PgW-IU4EX-M&feature=share [https://www.section.io/engineering-education/introduction-to-random-forest-inmachine-learning/](https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/)
3. [https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-ofdecision-tree-algorithm-428ebd199d9a](https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a)
4. Developing A Prediction Model For Customer Churn From Electronic Banking

Services Using Data Mining

Abbas Kermati, Hajar Ghaneei, and Seyed Mohammad Mirmohammadi

1. A Dynamic Classification Approach to Churn Prediction in Banking Industry
2. Hoiyin Christina Leung and Wingyan Chung
3. Machine Learning for Customer Churn Prediction in Retail Banking
4. Joana ias, Pedr Godinho, and Pedro Torres
5. <https://cran.r-project.org/web/packages/ROSE/ROSE.pdf>
6. <https://daviddalpiaz.github.io/r4sl/ensemble-methods.html>
7. <https://youtube.com/watch?v=tfN10IUX9Lo&feature=share>
8. https://youtu.be/JnlM4yLFNuo
9. [https://link-springer-com.stmarys-ca.idm.oclc.org/content/pdf/10.1007/s10586017-0933-1.pdf](https://link-springer-com.stmarys-ca.idm.oclc.org/content/pdf/10.1007/s10586-017-0933-1.pdf)
10. <https://europepmc.org/article/MED/25879060>
11. [https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-0180165-5](https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0165-5)
12. https://link.springer.com/content/pdf/10.1007%2F978-3-642-01973-9\_63.pdf
13. <https://machinelearningmastery.com/an-introduction-to-feature-selection/>
14. [http://www.sthda.com/english/wiki/cox-proportional-hazardsmodel#google\_vignette](http://www.sthda.com/english/wiki/cox-proportional-hazards-model#google_vignette)
15. [https://www.cs.bham.ac.uk/~pxt/NC/ncl.pdf(](https://www.cs.bham.ac.uk/~pxt/NC/ncl.pdf)NCL)
16. <https://arxiv.org/pdf/2011.13429.pdf>
17. <https://shiny.rstudio.com/articles/modules.html>(UI Shiny)
18. https://shiny.rstudio.com/articles/dynamic-ui.html
19. [Hands on Churn Prediction with R and comparison of Different Models for Churn Prediction](https://towardsdatascience.com/hands-on-churn-prediction-with-r-and-comparison-of-different-models-for-churn-prediction-4b79011a082a)
20. <https://towardsdatascience.com/predict-customer-churn-with-r-9e62357d47b4>
21. [http://rstudio-pubsstatic.s3.amazonaws.com/425842\_7f6b2293079f4a5d8fc7d62aeac2a545.html](http://rstudio-pubs-static.s3.amazonaws.com/425842_7f6b2293079f4a5d8fc7d62aeac2a545.html)
22. [https://towardsdatascience.com/project-modeling-predicting-of-churningcustomers-in-r-cb0a846ba94a](https://towardsdatascience.com/project-modeling-predicting-of-churning-customers-in-r-cb0a846ba94a)
23. https://www.shanelynn.ie/merge-join-dataframes-python-pandas-index-1/
24. <https://www.emilyzabor.com/tutorials/survival_analysis_in_r_tutorial.html>
25. https://www.datacamp.com/community/tutorials/survival-analysis-R

1. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ [↑](#footnote-ref-1)
2. https://youtube.com/watch?v=PgW-IU4EX-M&feature=share [↑](#footnote-ref-2)
3. https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-

   428ebd199d9a

   [↑](#footnote-ref-3)
4. https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ [↑](#footnote-ref-4)
5. *Developing A Prediction Model For Customer Churn From Electronic Banking Services Using Data Mining* [↑](#footnote-ref-5)
6. *A Dynamic Classification Approach to Churn Prediction in Banking Industry*

   Hoiyin Christina Leung and Wingyan Chung [↑](#footnote-ref-6)
7. https://cran.r-project.org/web/packages/ROSE/ROSE.pdf [↑](#footnote-ref-7)