
Customer Churn Predictions & Business Recommendations

Wealth Management Business 2nd Round Interview Case Study | March 2021

Philippe Heitzmann

Wealth Management Business 2nd Round Interview Case Study

Case Study Problem Statement

A key priority for the Wealth Management business is building and maintaining long-term relationships with WM clients. Senior leadership has asked you to develop an analytically based strategy to predict which WM clients are most likely to churn and how we can stem attrition. Please be prepared to present your findings and strategic recommendation in a one hour long panel interview to senior management within one week of receiving this prompt and dataset.

Dataset Overview

Variable Name	Description	Type
RowNumber	10K customers	Integer
CustomerID	Unique Identified for client	Integer
Surname	Client Surname	String
CreditScore	Ranging from 350 to 850	Integer
Geography	USWM Sales Division	String
Gender	Gender of customer	String
Age	Age of customer	Integer
Tenure	Length of client relationship in years	Integer
Balance	Investment balance snapshot	Float
Number of Products	Number of products	Integer
HasChckng	1 = Has a checking account; 0 = No checking account	Integer
IsActiveMember	1 = Digitally Active; 0 = Digitally Inactive	Integer
EstimatedSalary	Salary	Float
Exited	1 = Churned; 0 = Not Churned	Integer

Executive Summary

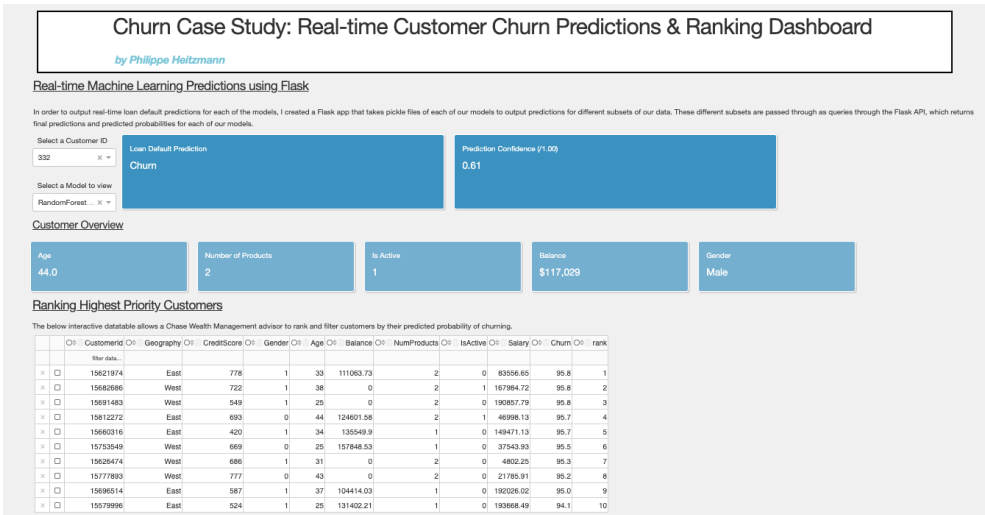
Overview

This case study analysis of customers churn rates in a US Wealth Management business puts forward a decision function parameterized by predicted probability of churning for a customer and business costs associated with churning to rank customers by greatest expected value of prescribing or not prescribing promotional activity.

Based on the feature importances of our models, we recommend that Wealth Management advisors providing customers with investment advice look to reinforce relationships with customers that are (i) older, (ii) using a relatively higher number of products, (iii) are active members, (iv) have high balances and (v) identify as female. Assignment to the East USWM division being an important predictor in all our models may indicate (i) an emerging competitive threat in the East region that should be addressed or (ii) the need to investigate the East USWM division’s management practices in the case poor management is leading to increased customer attrition.

We finally create a [dashboard experience](#) using Dash and Flask Python web frameworks in order to make our customer ranking results available to Wealth Management advisors to better inform targeted marketing and sales outreach at the single-customer level.

Dash & Flask App Dashboard View



Recommended Customer Subsets to Target

We recommend targeting customers in the following subgroups:






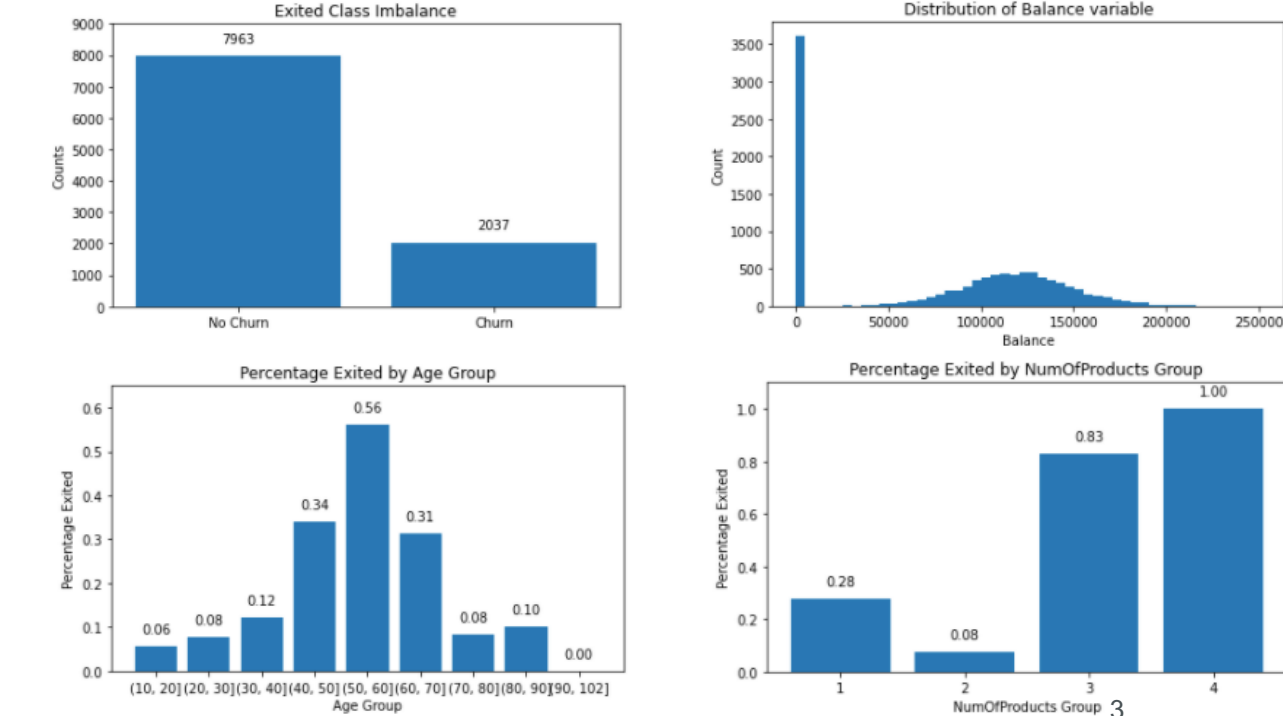
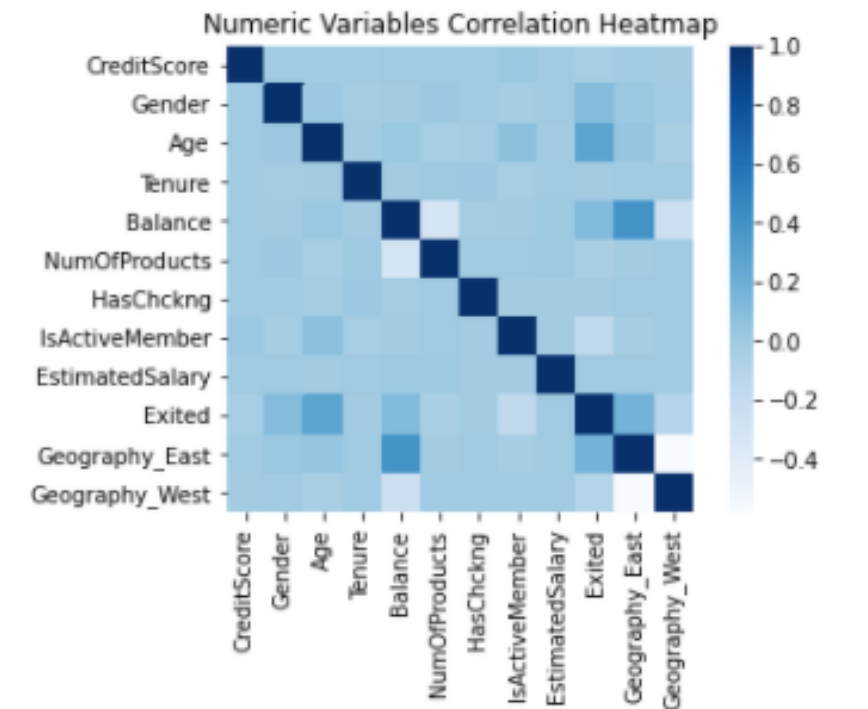
-  1 >40 years old
-  2 Using >2 Products
-  3 Count as Active
-  4 Account balances >\$75k
-  5 Identifying as Female

Table of Contents

- ① **Exploratory Data Analysis**
- ② **Data Preprocessing**
- ③ **Feature Engineering**
- ④ **Machine Learning Classifications**
- ⑤ **Business Recommendations**
- ⑥ **Dashboarding Results**

- 100





Data Preprocessing – Dealing with Outliers and Missing Data

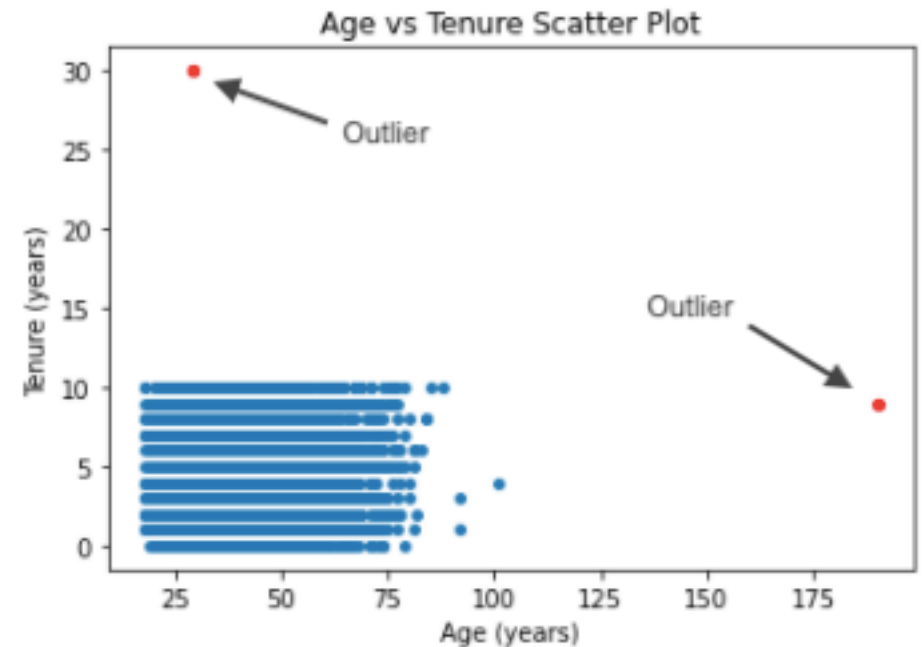
Overview

We identify two impossible outlier values in the Age and Tenure variables that we attribute to data entry errors and would skew the results of our linear models if left unaddressed and impute these using linear regression and mean imputation respectively. Three missing CreditScore values are imputed to the mean based on domain knowledge and analysis of other data on these customers while two CreditScore values of 305 & 865 are kept per outside research that CreditScores occasionally overshoot the 350-850 range specified in the provided data dictionary

Identifying & Treating Outliers in our Dataset

- We notice two impossible outlier values of 190 in the Age variable and 30 in the Tenure variable when Age is 29 for that customer
- Given the relatively high correlation value of 0.08 between Age and IsActiveMember, we choose to impute this Age of 190 value using simple linear regression trained on IsActiveMember and impute this value to 37.8
- Given the previous correlation heatmap did not show any particularly strong linear relationships between Tenure and the other x variables in our dataset, we choose to impute this incorrect value by the mean to 5.0
- While our three missing CreditScore values initially stand out as all belonging to Male customers above the age of 35 not from the East Region, we observe that these are all customers with salaries >\$120k, using at least one banking product, with two having non-insignificant savings balances of >\$80k, which tells us these are customers that should realistically not only have a CreditScore but have a CreditScore that is not too low
- Given the previous correlation heatmap did not show any particularly strong linear relationships between CreditScore and the other x variables in our dataset, we choose to impute this incorrect value by the mean to 650
- One CreditScore value of 865 is left untreated per outside research that FICO CreditScores can occasionally surpass 850 for certain customers despite data dictionary indicating this variable ranges from 350 to 850

Visualizing Age & Tenure Outliers



Feature Engineering – Extracting Demographic Information from Surnames from Census data

Overview

2010 Census Surname Data provides information on frequency of 162,255 Surnames in the United States and offers demographic information on ethnicity percentages by Surname. We left join our dataset with the census dataset on Surname in order to extract this average ethnicity breakdown by Surname.

Census 2010 Surname Dataset

Variable Name	Description	Type
Name:	Surname	string
Pct2Prace:	Percent Non-Hispanic Two or More Races	int
Pctaian:	Percent Non-Hispanic American Indian and Alaska Native Alone	int
Pctapi:	Percent Non-Hispanic Asian and Native Hawaiian and Other Pacific Islander Alone	int
Pctblack:	Percent Non-Hispanic Black or African American Alone	int
Pcthispanic:	Percent Hispanic or Latino origin	int
Pctwhite:	Percent Non-Hispanic White Alone	int
Prop100K:	Proportion per 100,000 population for name	int
Rank:	National Rank	int
Count:	Frequency: number of occurrences nationally	int
Cum_Prop100K:	Cumulative proportion per 100,000 population	int

- After cleaning the Name variable in the Census dataset and matching capitalization formats we perform a left join of our dataset on the Census dataset on the Surname variable
- Missing demographic variable values of Surnames in our dataset not captured in the Census dataset are imputed using the percentage frequency of that ethnic subgroup in the general population per Census data

Feature Engineering – Extracting Information about Potential Relatives from Surnames from Census data

Overview

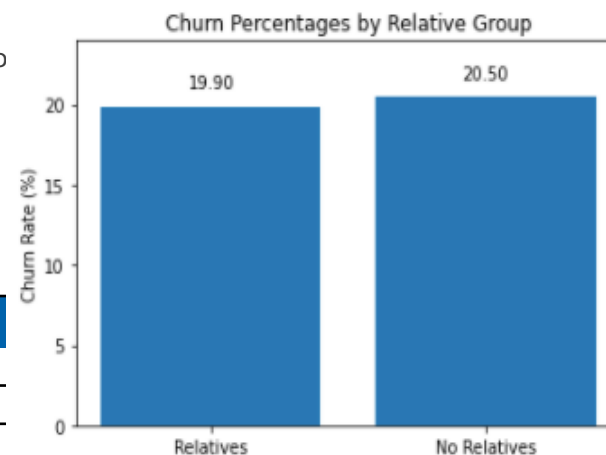
As our Census data includes information on Surname frequency per 100,000 for a given name, we create a new variable 'Relative' capturing whether or not we think with a reasonable degree of certainty that two or more customers with the same uncommon last name are related. Our hypothesis is that Customers that might be related based on this measure may exhibit group tendencies to churn at higher or lower rates than customers with no customer relatives

Creating *Relative* variable from Surname

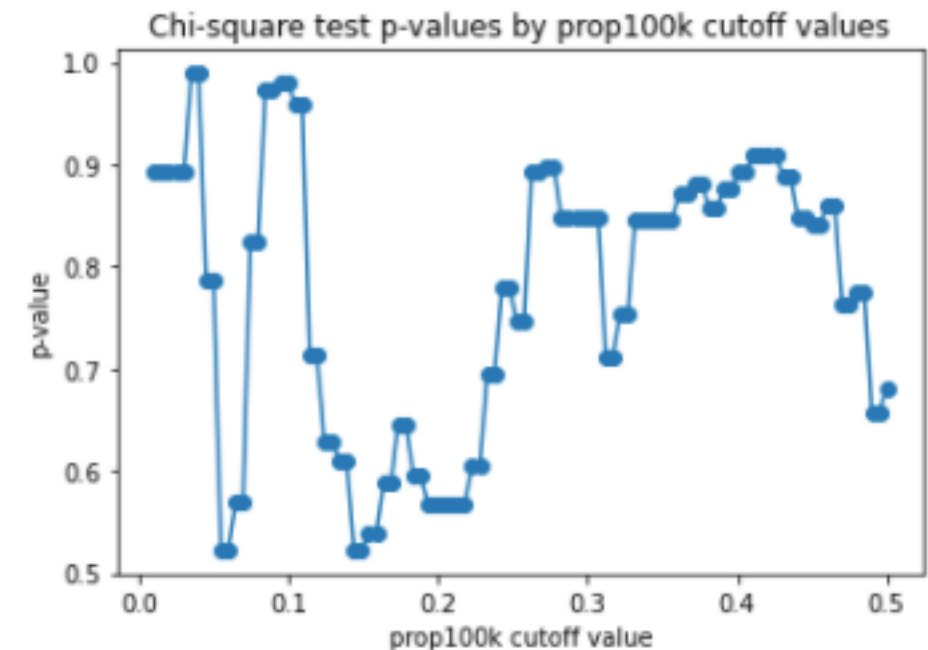
- We use the Census *prop100k* variable to decide if a given Surname should be considered common or uncommon
- In order to pick the optimal *prop100k* cutoff value to designate groups of two or more repeat Surnames in our dataset as potential relatives, we perform multiple Chi-square tests of dependence between the frequencies of *Exited* for groups of two or more customers with identical Surnames and *prop100k* frequencies below a certain threshold using 50 evenly spaced cutoff values in the interval [0.01, 0.5] and record the p-value statistic for each cutoff
- While our chi-square tests p-values show that we cannot reject the null hypothesis that our *Relative* and *Exited* variables are independent, we choose to still use the lower cutoff value of 0.055 to create a *Relative* variable where we think that person may have with a reasonable degree of certainty a relative in the dataset

Table 1. Exited Frequencies for *prop100k* cutoff value of 0.055

	Not_Related (=0)	Related (=1)
Exited	1540	497
Did_Not_Exit	5963	2000



Testing Hypothesis with Chi-square test of dependence



Model Training & Performance across LogisticRegression, RandomForestClassifier & XGBoostClassifier

Overview

We choose to train `LogisticRegression`, `RandomForestClassifier` and `XGBoostClassifier` models in order to test the relative strengths and weaknesses of linear, horizontal tree-ensembling and vertical tree-ensembling / boosting frameworks respectively when applied to our dataset.

Model Performances

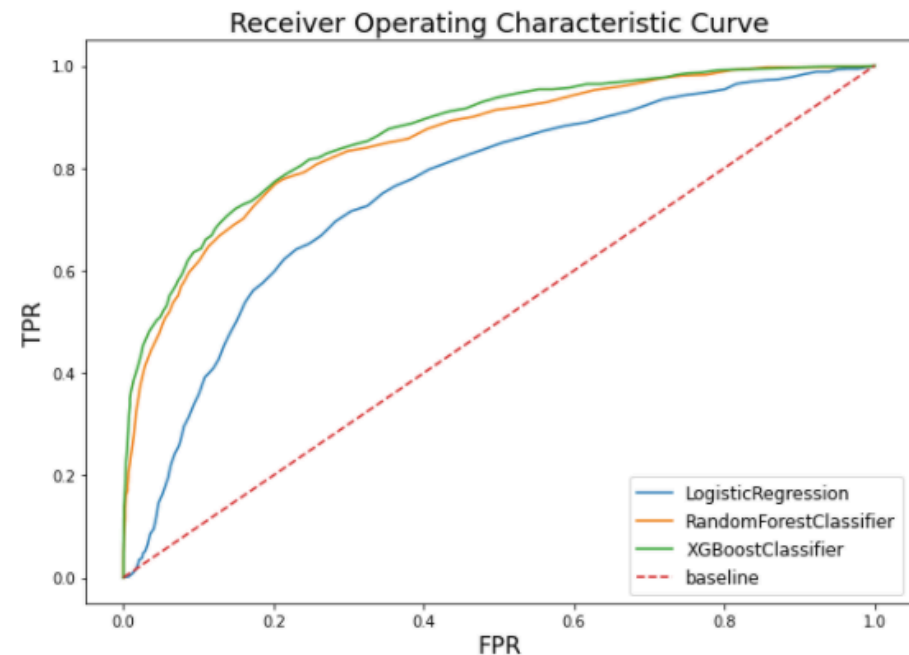
- For purposes of choosing a performance metric to train and tune our models, we assume that the business will incur significantly higher costs from customers who churn and were not prescribed a treatment than from customers who do not churn and are prescribed a treatment, which we define as some type of promotional action such as discounted management fees to incentivize the customer to not churn
- Given this relatively higher costs of False Negatives we choose recall score as the most relevant metric to compare model performances and tune our models to optimize recall score

Table 2: Cross-Validated ROC_AUC, Precision & Recall Scores for Trained Models

	LogisticRegression	RandomForestClassifier	XGBoostClassifier
ROC_AUC	0.7482 ($\sigma = 0.022$)	0.8553 ($\sigma = 0.024$)	0.8707 ($\sigma = 0.025$)
Precision	0.6447 ($\sigma = 0.021$)	0.4956 ($\sigma = 0.032$)	0.5000 ($\sigma = 0.023$)
Recall	0.7097 ($\sigma = 0.039$)	0.7560 ($\sigma = 0.047$)	0.7693 ($\sigma = 0.031$)

- We adjust for class imbalance when training our models and stratify our `train_test_splits` in order to ensure each split sees equivalent proportions of 1's and 0's in each of our sets
- Of all our models, `XGBoostClassifier` provides the highest recall score with the lowest standard deviation as well as highest ROC_AUC score
- `XGBoostClassifier predict_proba_` probabilities will therefore be used in the next stage to drive our business recommendations and strategy to deal with churn

ROC_AUC Curves by Model



Addressing Predicted Churn with Business Promotional Strategy

Overview

We define a business promotional strategy leveraging our XGBoostClassifier model predicted churn probabilities that takes into account cost of a customer churning, cost of promotional activity prescribed to address potential churn, anticipated effectiveness of promotional activity in stemming churning, and cost of prescribing the wrong action in order to rank customers by greatest expected value of prescribing or not prescribing promotional activity

Simple Rule-Based Business Strategy to Address Potential Churn

- We create a function parameterized by predicted probabilities of churning for a customer, cost of a customer churning, cost of promotional activity prescribed to address potential churn, anticipated effectiveness of promotional activity in stemming churning, and cost of prescribing the wrong action to produce a simple business strategy to address or not to address expected churn for a single customer
- This function lets the user select different values for these inputs depending on the cost of each in order to (i) yield a subset of customers that should be the target of promotional activity and (ii) calculate an expected value in dollars from the prescribed promotional activity in order to allow for ranking and prioritization of customers in the case that business resources are constrained and not all customers can be prescribed promotional activity
- In this case we leverage our XGBoostClassifier predicted probabilities given this model produced the highest recall score of all our models in order to return this customer subset that should be targeted with promotional activity
- Our model recommends taking promotional action on 1,510 customers in our dataset based on inputting the following values for our function arguments, down from the ~2,500 customers that were predicted as churning by our XGBoostClassifier model based on a decision cutoff value of 0.20

Table 3: Function Arguments to recommend taking promotional action for 1,510 customers

Function Parameter	Average Dollar Cost
Churn Cost	\$1000
Promotion Cost	\$500
Promotion Effectiveness (/100%)	80%
Cost of Wrong Action	\$1

Table 4: Head of Returned Customer Subset ranking customers recommended as targets for promotional activity by expected value of promotion based on Table 3 attributes

Customer Index	Pred Churn Prob	Target with Promotion?	EV of Promotion
1258	95.84%	Yes	\$265.67
2578	95.83%	Yes	\$265.65
645	95.82%	Yes	\$265.59
1208	95.74%	Yes	\$264.95

Client Relationship Management Recommendations based on Model Feature Importances

Overview

Based on feature importances of our models, we recommend that Wealth Management advisors providing customers with financial planning services look to reinforce relationships with customers that are (i) older, (ii) using a relatively higher number of products, (iii) are active members, (iv) have high balances and (v) identify as female. Assignment to the East USWM division being an important predictor in all our models may further indicate the need to investigate the East USWM division’s management practices in the case poor management is leading to increased customer attrition.

Client Relationship Management Recommendations


- Observing that Age is the #1 ranked predictor across all models, we recommend that advisors develop strategies for creating greater customer satisfaction for older customers in the >40 age group and expand print marketing outreach to these customers
- We further recommend that advisors, if possible, seek to collect survey information from active members with high balances using 3 or 4 products to see if potential user experience dissatisfaction when simultaneously managing investment portfolios across these different digital platforms is causing higher churn rates in this subgroup
- We recommend that advisors look to implement customer retention strategies targeting female-identifying customers given this subgroup is 1.5 times more likely to churn than their male-identifying counterparts
- A customer being assigned to the East USWM division being a top five ranked feature across all models appears to indicate that structural factors, such as poor management, may be leading to higher customer attrition in this division, or that a competitive threat from a regional Wealth Management player in the East region may be stealing customers, and should be investigated further
- We further note that our Relative variable created earlier through feature engineering leveraging our Census data is ranked near the bottom in terms of feature importance in each of our models indicating this variable should not be of high focus for future analyses


Feature Importances Rankings by Model

Table 5: Feature Importances Rankings for XGBoostClassifier, RandomForestClassifier & LogisticRegression

	XGBoostClassifier	RandomForestClassifier	LogisticRegression
Age	1	1	1
NumOfProducts	2	2	10
IsActiveMember	3	3	14
Balance	5	5	4
Gender	7	6	3
Geography_East	4	4	2

Recommended Customer Subsets to Target

① >40 years old

③ Are Active members

⑤ Female

② Using > 2 Products

④ Have account balances >\$75k

Building a DashApp Web Experience leveraging a FlaskApp Backend to Dashboard Results for advisors

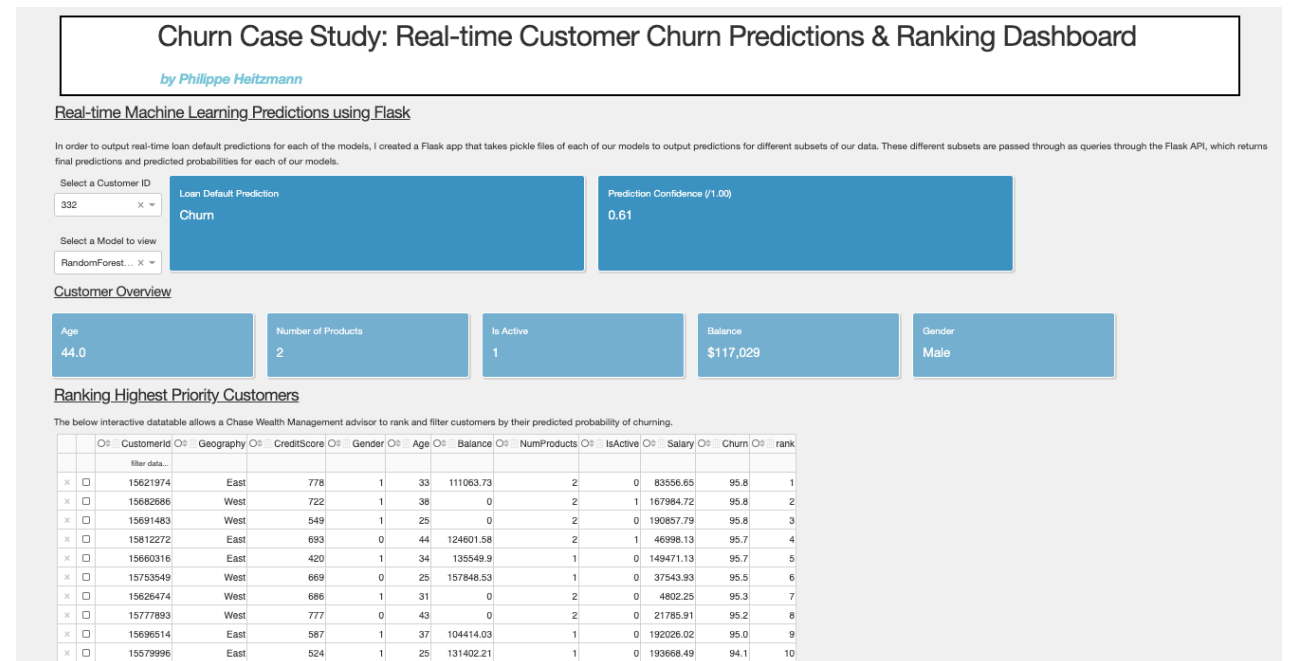
Overview

We build a [dashboard experience](#) using Dash and Flask Python web frameworks to make our customer ranking results available to Wealth Management advisors in order to better inform targeted marketing and sales outreach to individual customers. This Dash application is further productionalized using an Amazon Web Services Elastic Compute instance and is accessible to any web traffic at <http://ec2-3-17-150-87.us-east-2.compute.amazonaws.com/>.

Client Relationship Management Recommendations

- Our dashboard allows advisors to filter a provided table of customers by CustomerId, predicted churn probability and customer prioritization rank in order to quickly find out which customers should be prioritized and targeted for promotional sales outreach
- This Dashboard further allows advisors to select a CustomerId in a dropdown in order to get real-time churn predictions through serialized versions of our LogisticRegression and RandomForestClassifier models loaded into our Flask app backend
- Our models are passed x variables of that CustomerId athrough the REST API Flask endpoint and ping back prediction probabilities to our Dash frontend in JSON format
- The dashboard further updates Age, NumberOfProducts, IsActiveMember, Balance and Gender information on the selected CustomerId in order for the advisor to build business intuition around some of the factors that may lead to higher predicted customer churn rates

Dash App View



Additional Candidate Information

Other Portfolio Projects

For reference please refer to my blog and Github at the below links for an overview of some of the other data science projects I have been involved with.

Blog:

<https://nycdatasience.com/blog/author/philippe-heimann/>

Github:

<https://github.com/philippe-heimann>