S470/670 - Final project: Client Churn Analysis

(He / Him) Sreesha Srinivasan Kuruvadii - sskuruva@iu.edu
(She / Her) Asha Pondicherry - asrpondi@iu.edu
(She / Her) Himani Shah - shahhi@iu.edu

Introduction

The US Wealth Management (USWM) business unit comprises Chase Wealth Management, JP Morgan Securities, and the You Invest digital platforms. Chase Wealth Management (CWM) is a branch-based model in which a dedicated Advisor offers comprehensive financial planning and investment management advice to individuals and families. A key priority for Chase Wealth Management is to develop an analytically based strategy to check which CWM clients are most likely to churn which is also known as attrition or defection, refers to the loss of customers from a company's customer base.

Dataset

We have a dataset of 10000 rows with 14 attributes to perform our analysis and give recommendations to stem attrition. (Source: JPMC take home competition data)

Categorical Features	
Gender	Sex of the client [Male, Female, Neutral]
Geography	sales division [Central, East, West]
HasChecking	1: has a checking account, 0: otherwise
IsActiveMember	1: Digitally Active, 0: Digitally Inactive
Numerical Features	
Age	Age of the client [years]
Credit Score	Ranging from 350 to 850
Tenure	Length of client relationship in years
Balance	Account balance snapshot
Number of Products	The number of products a client is associated
Estimated Salary	Salary of the clients
Target Variable	
Exited	1: client churned, 0: client did not churn

Churn

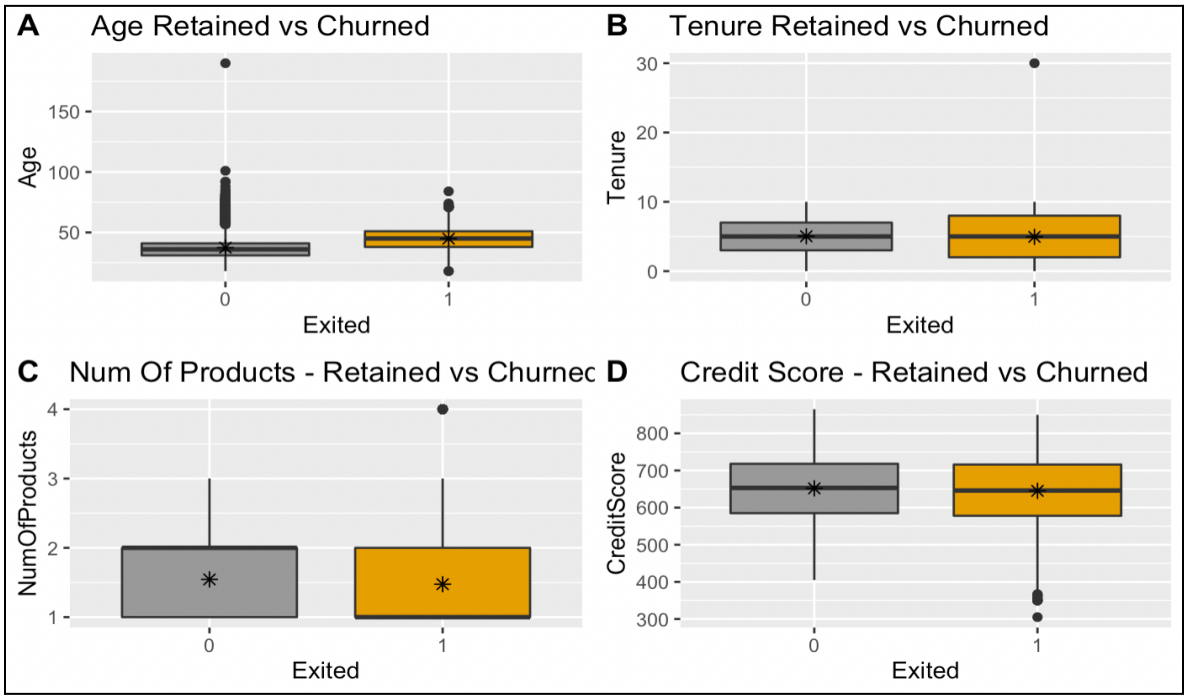
When a client stops using products of JPMC they are considered to be churned. This is represented with Exited (Target Variable).

Research Questions

We try to address this issue by performing exploratory analysis and answering the following questions:

1. Are all features important for determining whether the client will churn or not?
2. How do various features impact the likelihood of churn?
3. Could we well describe the notions of churn using a simple logistic regression or do we need more complex models?

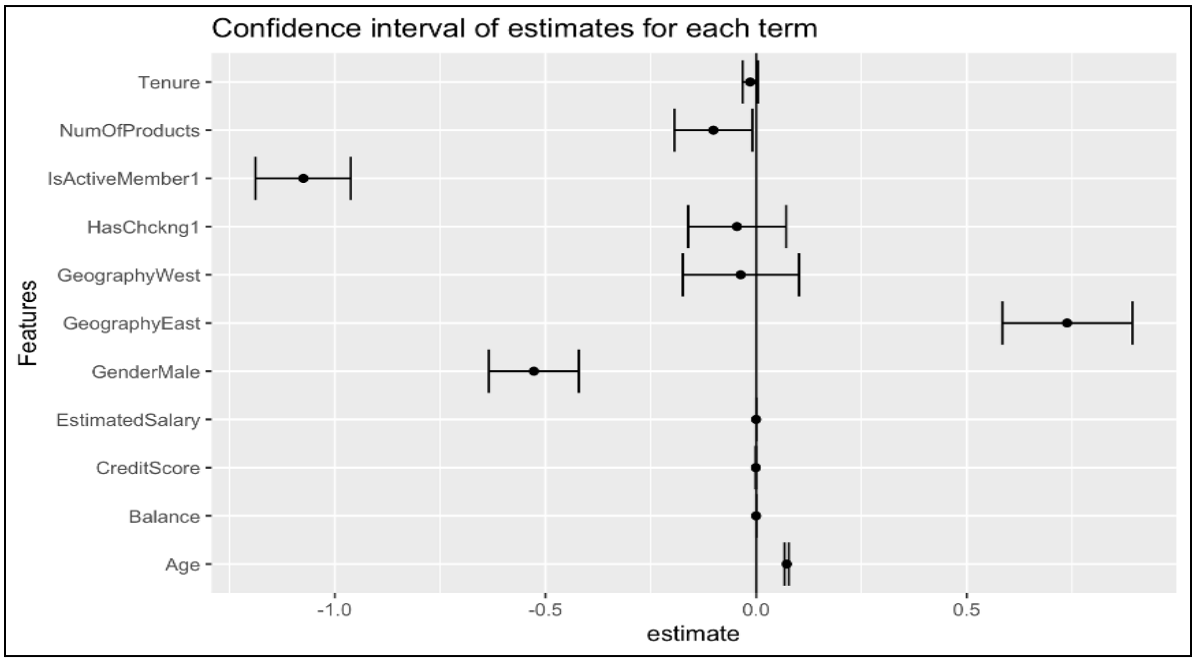
Data Inspection and cleaning:



Removed/imputed such data from our analysis:

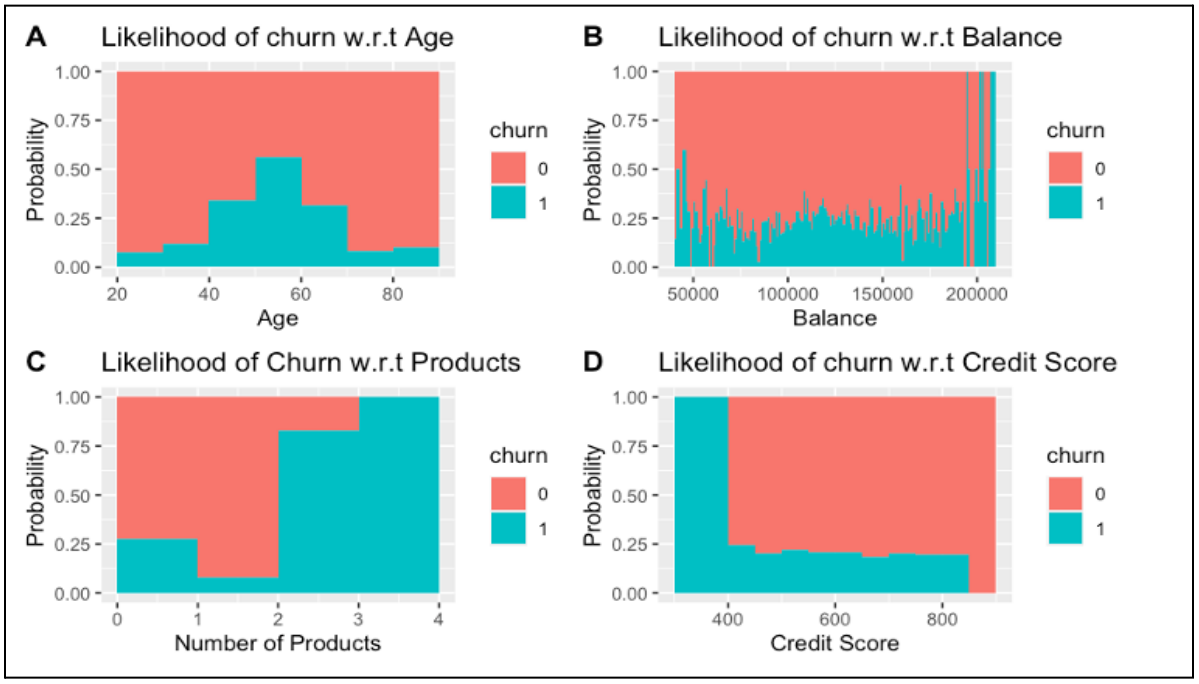
- We can see some noise in the Age column for the clients who are retained. That is age cannot be 190, there is one other client with age 101.
- There is literally one client with 30y tenure and the is age is 29yrs
- There is a small group of clients with 4 number of products who are churned.
- The average salary of the clients who are churned and retained are the same and we do not see any significant effect on the likelihood of churn w.r.t clients salary. Usually there should be an effect on churn for example - clients with lower salary are likely to churn because they cannot afford to invest. But we do not see such a scenario with given data.

Important Features



Create a logistic regression model with all the predictors and observe the results. A lot of the coefficients are within the margin of error of zero. Some features can be eliminated. Fewer predictors would perform better. According to the above plot we see IsActiveMember, Tenure, NumOfProducts, Geography, Gender, Estimated Salary, Age are the important features and we use the same in our models going further.

Numerical Features as Explanatory Variables



Using Age as an explanatory variable, we observe that the likelihood of client churn increases between 40 and 70 age groups. The likelihood of client churn increases with an increase in the number of products. Clients with lower credit scores are more likely to churn. We see there is a lot going on with Balance which will be investigated in our further analysis.

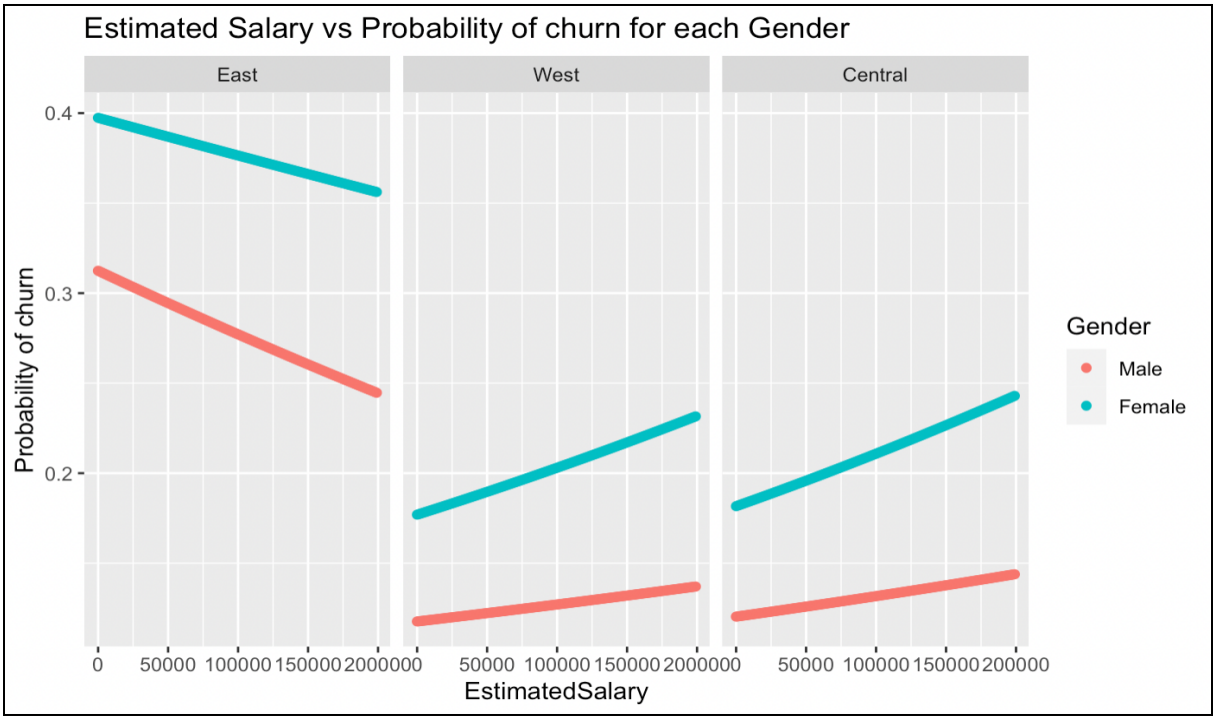
Two Predictor Model

Transformation of Number of Products decreases the AIC significantly.

```
gam(Exited ~ s(Age, bs = "cr") + s(Balance) + log.NOP * IsActiveMember + Gender + Gender:Geography, family = binomial(link="logit"), data =data)
```

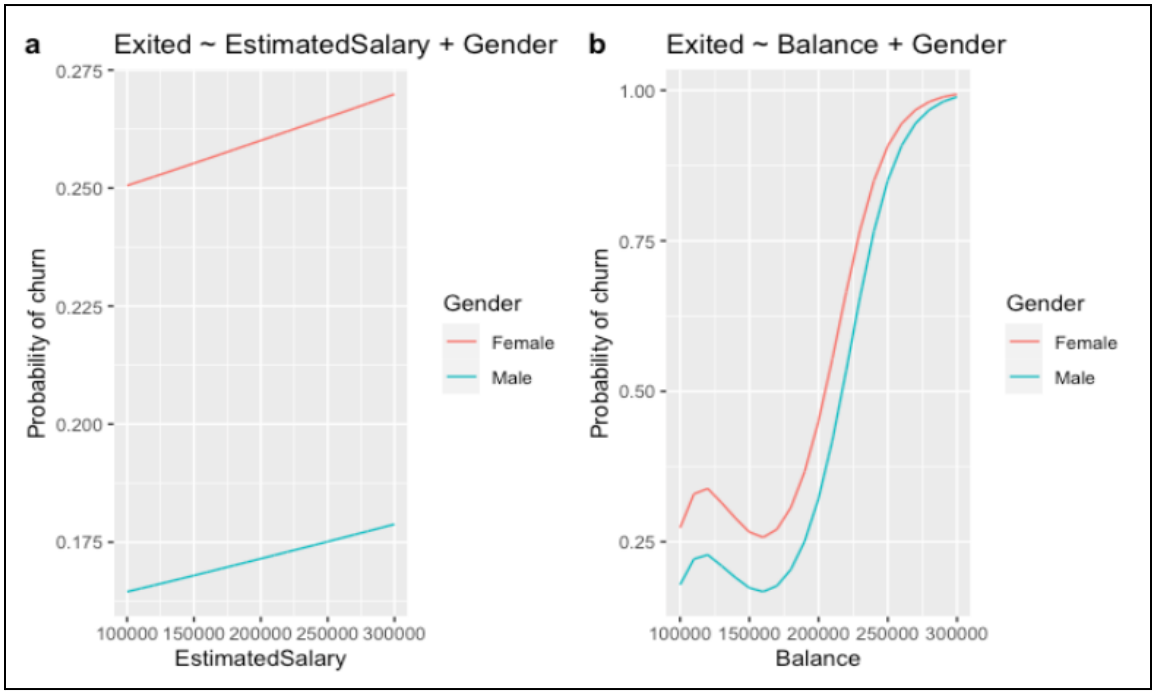
logit[P(Churn| Age, NumOfProducts)] = -3.652 + 0.06 * Age - 0.177 * NumOfProducts

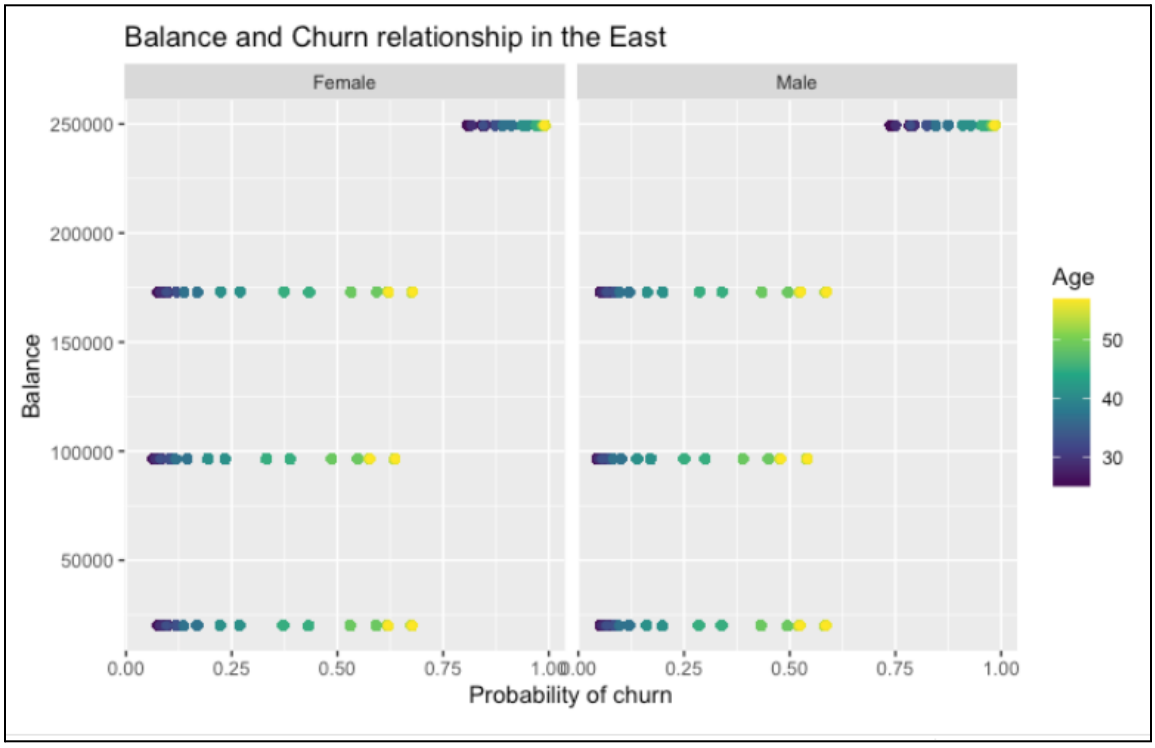
Probability of Churn with Interactions: Estimated Salary, Balance and Gender



In the above plot we see that Female clients have more probability of churn across all the three regions, with the increase in salary churn rate also increases but in Eastern region we notice that initially the churn rate is high and it monotonously decreases both in male and female clients. These are our findings with GLM, let's try out GAM and see if we can make different conclusions.

Balance and Estimated Salary impact on churn



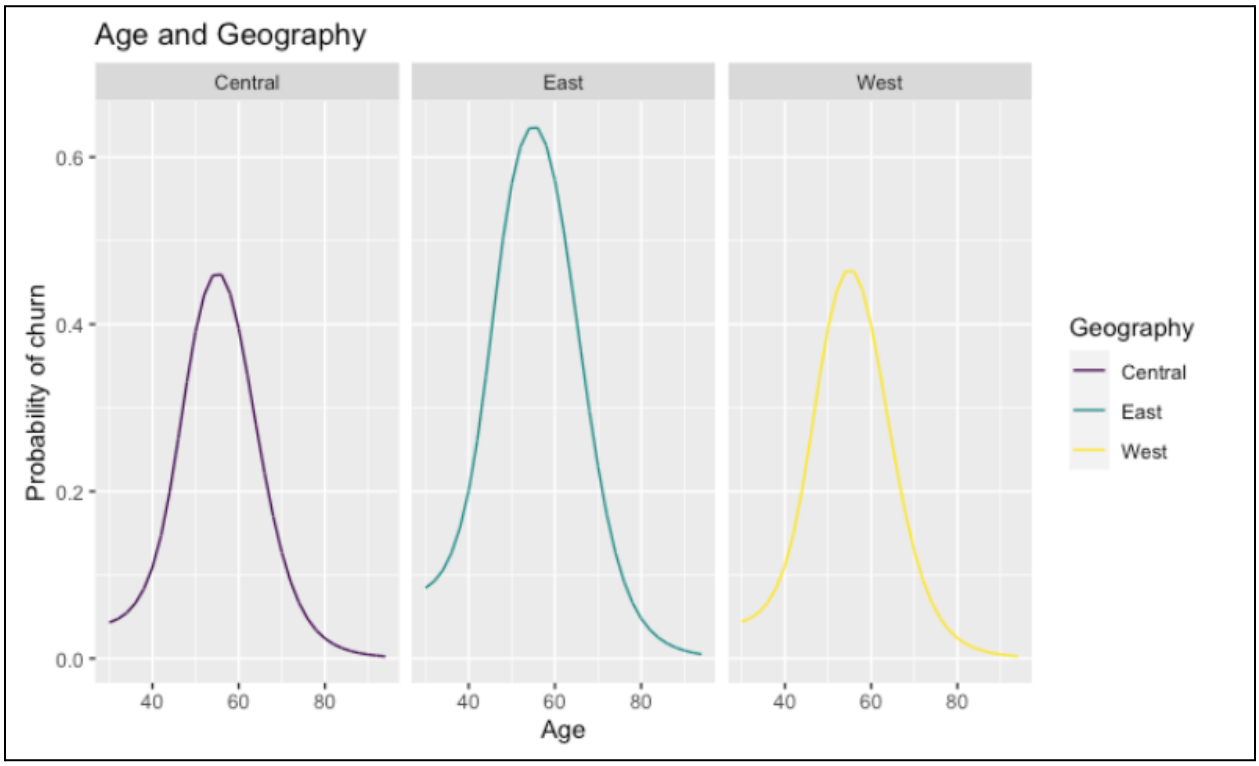


Churn is highest at a Balance of 250000. We cannot conclude anything significant since the churn across balance ranges are similar and abruptly increases at 250000. But Age groups around 25 - 35 have the lowest churn probability and age groups post 50 have the highest churn probability.

Categorical Features as Explanatory Variables:

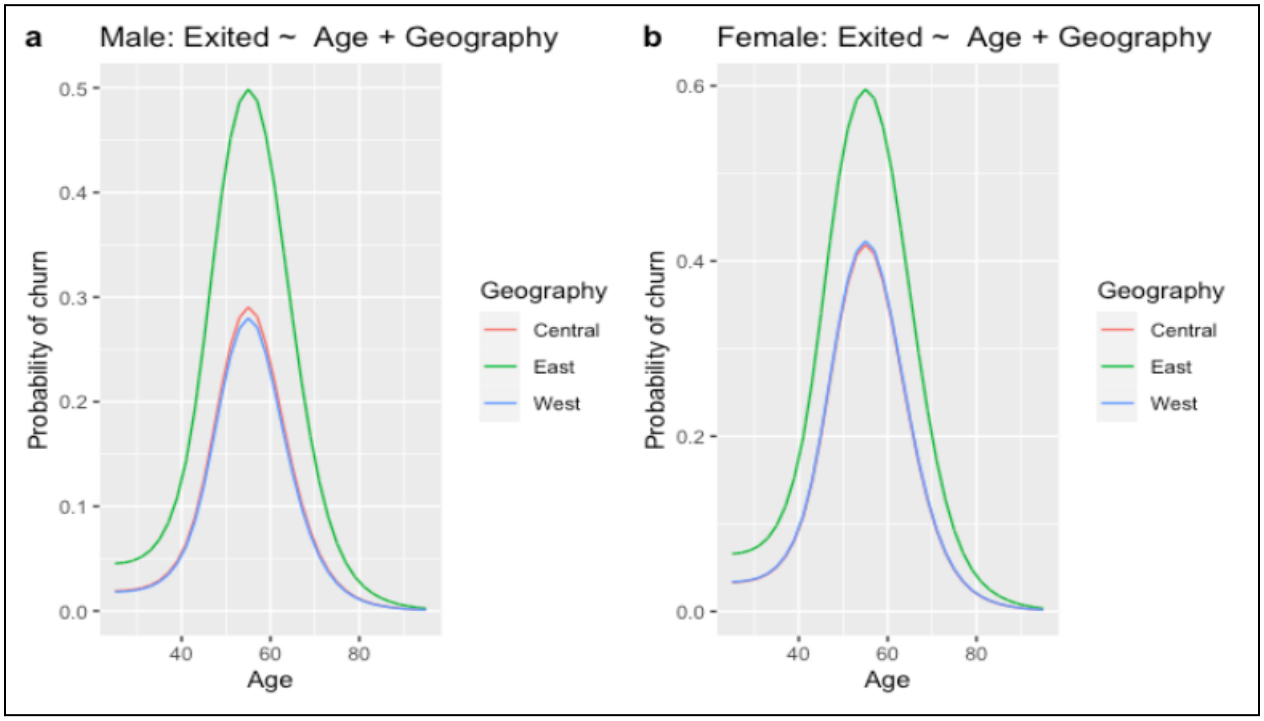
Analysis of Geography

People of age 50 have the highest churn rate. People above 75 and below 25 show the lowest tendency to move out of the system. People in the east are more prone to churn than West and Central. Churn is high for age groups between 30 and 65.



Geographically, in the East – an increase in the salary earned increases the tendency for the customer to exit much more than a similar increase in their account balance. Eastern regions have a higher tendency to churn, but the difference is quite insignificant and may not add much value. We need more data , possibly across continents to start a stronger analysis.

Probability of churn with each value of Gender



Females have the highest tendency to exit from the system. As a common trend both male and females around the age of 50 have a higher tendency to exit the system.

GLM vs GAM

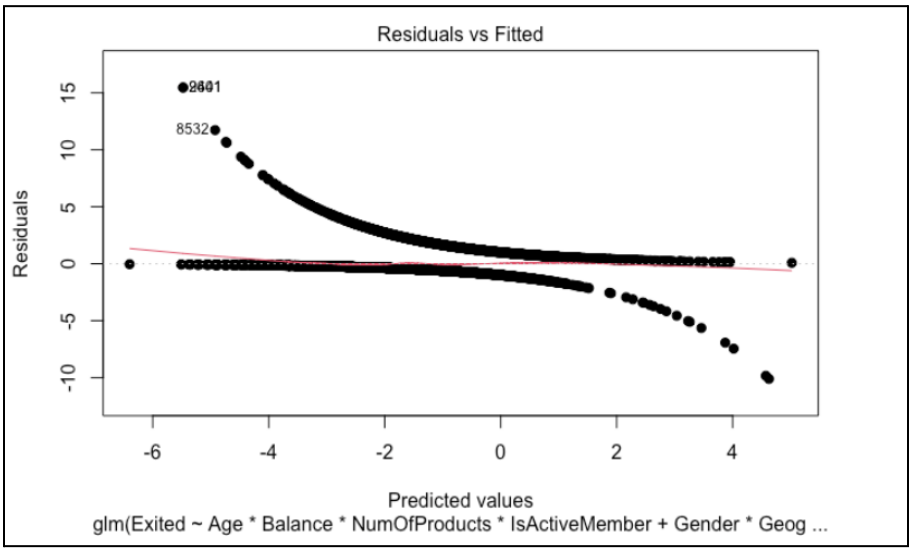
```r
AIC(gam(Exited ~ s(ScaledAge, bs = "cr") + s(ScaledBalance) + ScaledTenure:ScaledCreditScore, family =
binomial(link="logit"), data =data))
AIC(gam(Exited ~ s(Age, bs = "cr") + s(Balance) , family = binomial(link="logit"), data =data))
AIC(gam(Exited ~ s(Age, bs = "cr") + s(Balance) + log.NOP * IsActiveMember + Gender + Gender:Geography+
ScaledTenure:ScaledCreditScore, family = binomial(link="logit"), data =data))

AIC(gam(Exited ~ s(Age, bs = "cr") + s(Balance) + log.NOP * IsActiveMember + Gender + Gender:Geography,
family = binomial(link="logit"), data =data))
```

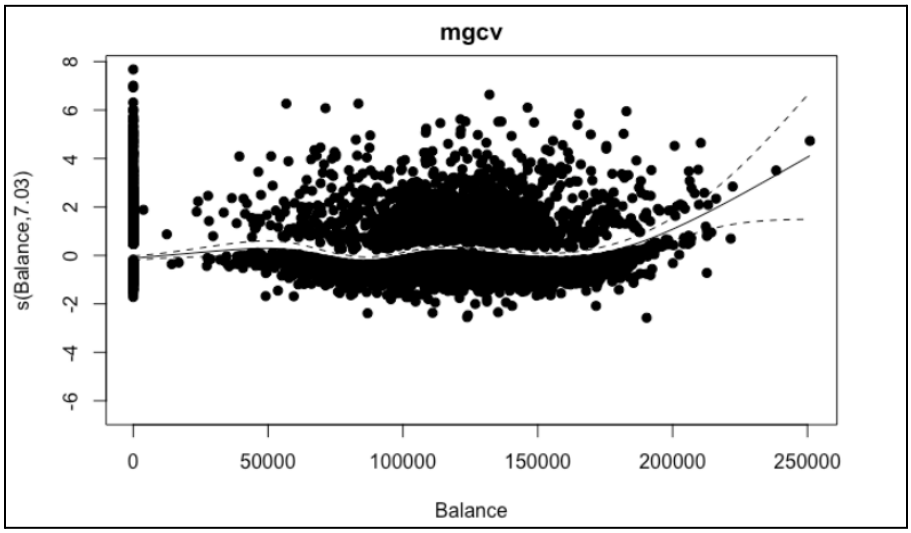
```
[1] 8533.559
[1] 8533.564
[1] 7965.035
[1] 7964.761
```

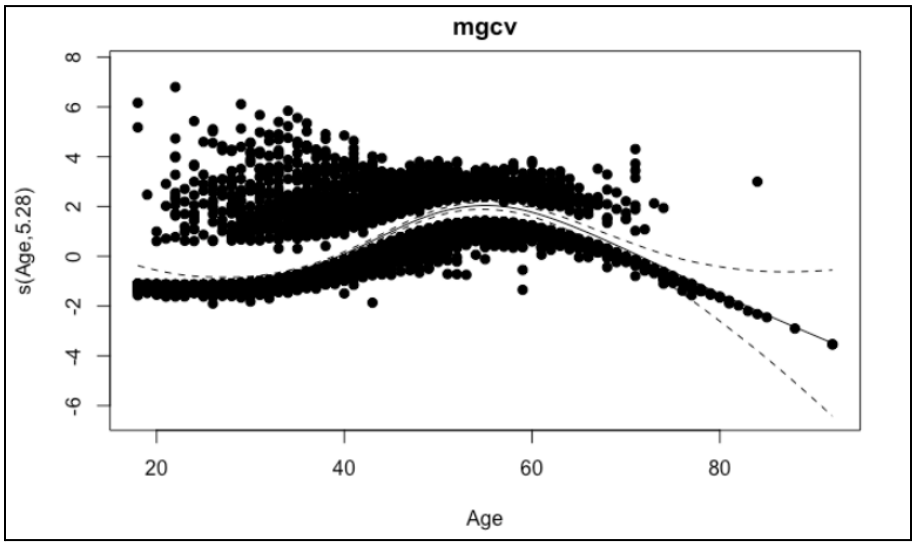
GAM has the lowest AIC score. We chose GAM for this reason.

Residuals with GLM



Residuals with GAM





```
glm(Exited ~ Age * Balance * NumOfProducts * IsActiveMember + Gender * Geography ,
family = binomial(link="logit"), data =data)
```

```
gam(Exited ~ s(Age, bs = "cr") + s(Balance) + log(NumOfProducts) * IsActiveMember +
Gender * Geography + Tenure, family = binomial(link="logit"), data =data)
```

GAM has a significant lower AIC value than GLM, meaning it represents data better. Therefore we go by GAM for our modeling choice. This correlates with the plots since GLMs have residuals with trends and GAMs have residuals trending around zero. GAM also shows a non-linear non-monotonous trend for Age, whereas GLM fails to capture this.

Conclusions

Females have a higher tendency to churn out than males. An increase in balance has an influence on the probability of churn. East always has a higher churn probability than other regions. But this is quite smalle around ~0.1 - 0.2 difference. We do not have enough data to investigate further. GAM works better than GLM to model the interactions between Number of Products used, IsActiveMember, Gender and Geography, and Churn. Fewer predictors would perform better than using all the features together. We looked at how different features affect the likelihood of Churn by plotting them with our target variable and fitting a Logistic regression model using all features and a subset of them. Taking a subset of features to fit the model gives a lower AIC score.

Limitations & Future Work

Our Primary Limitation is that we lack data for multiple regions across continents. Within the USA there seems to be very little change in the churn. Even though the mean residual is centered around zero, the range in residual values is quite high suggesting some kind of transformation is required. We would want to try tree-based models/boost models to capture complex interactions between numerical and categorical variables.