Capstone Project: Cost of Churn Prediction

Tony Cronin

4/25/2021

Executive Summary

What is customer Churn?

Customer Churn is defined as loosing customers leaving your business or service for one of your competitors.

- (Hadden et al. 2007) tell us "a business incurs much higher charges when attempting to win new customers than to retain existing ones."
- (Hadden et al. 2007) tells us "companies have acknowledged that their business strategies should focus on identifying those customers who are likely to churn."
- According to (Frank Sherlock 2018) churn costs Business Billions. In one example, he estimates churn costs energy providers "£25.05 billion per annum" in the UK market alone.
- Stripling suggests (Stripling et al. 2018) in saturated markets acquiring new customers is "eminently challenging, and costs five to six times more than to prevent existing customers from churning"
- We can use ML to predict customer churn, but there is a problem, (Lemmens and Gupta 2020) state "conventional approach has been to target customers either based on their predicted churn probability" which does not reflect business value. In Machine Learning we evaluate the cost of a model with AUC or RMSE scores, however in a business context not all predictions are costed the same way (Lemmens and Gupta 2020) state that standard ML predictions "ignore that some customers contribute more to the profitability of retention campaigns than others."

Conclusion: businesses that develop strategies to reduce customer could increase profits and block the growth of their competitors, but any strategy employed must pay close attention to the cost of our prediction.

Analysis

There is very little data on customer churn available in the public Domain. For most business' churn data is a closely guarded secret. One of few available is **Bank Customer churn Data**, which was released in 2019.

We will investigate the churn predictors in this data set.

Understanding our data

In **Bank Customer Churn**:

- we have 10000 observations and 11 variables
- Variables are: Churn, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary.

```
glimpse(bank_churn_data_tbl)
```

```
## Rows: 10,000
## Columns: 11
## $ Churn           <dbl> 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ~
## $ CreditScore     <dbl> 619, 608, 502, 699, 850, 645, 822, 376, 501, 684, 528,~
## $ Geography       <fct> France, Spain, France, France, Spain, Spain, France, G~
## $ Gender          <fct> Female, Female, Female, Female, Female, Male, Male, Fe~
## $ Age             <dbl> 42, 41, 42, 39, 43, 44, 50, 29, 44, 27, 31, 24, 34, 25~
## $ Tenure          <dbl> 2, 1, 8, 1, 2, 8, 7, 4, 4, 2, 6, 3, 10, 5, 7, 3, 1, 9,~
## $ Balance         <dbl> 0.00, 83807.86, 159660.80, 0.00, 125510.82, 113755.78,~
## $ NumOfProducts   <dbl> 1, 1, 3, 2, 1, 2, 2, 4, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, ~
## $ HasCrCard       <dbl> 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, ~
## $ IsActiveMember  <dbl> 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, ~
## $ EstimatedSalary <dbl> 101348.88, 112542.58, 113931.57, 93826.63, 79084.10, 1~
```

```
summary(bank_churn_data_tbl)
```

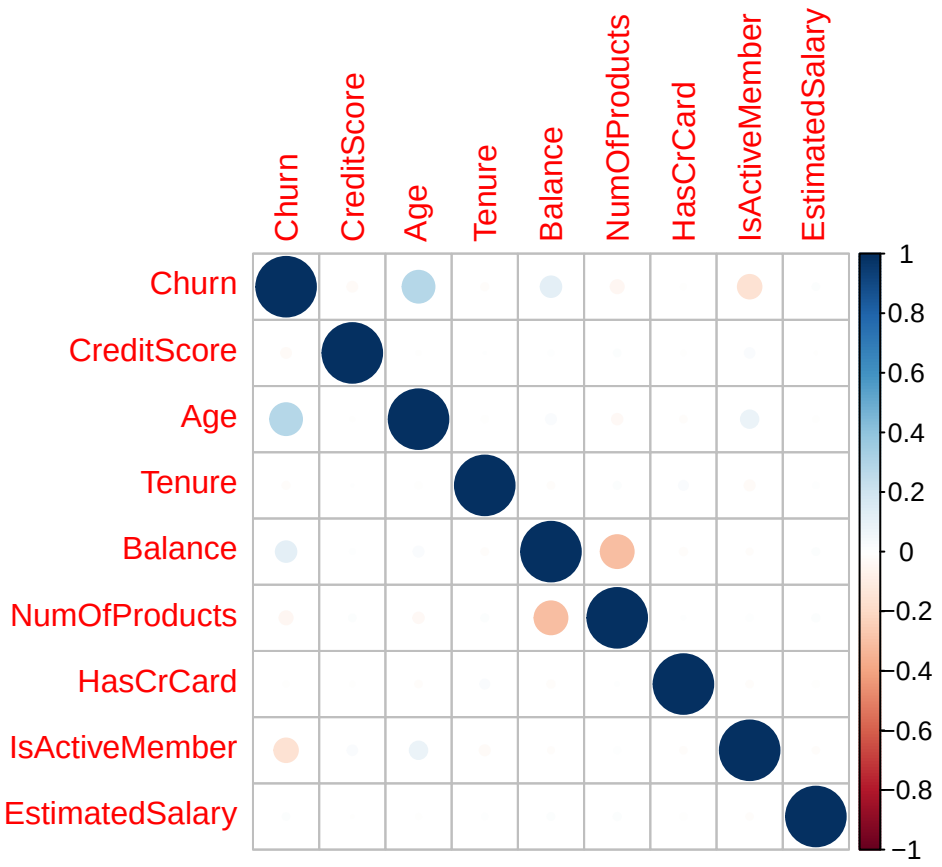
```
##      Churn         CreditScore      Geography       Gender          Age
##  Min.   :0.0000   Min.   :350.0   France :5014   Female:4543   Min.   :18.00
##  1st Qu.:0.0000   1st Qu.:584.0   Spain  :2477   Male  :5457   1st Qu.:32.00
##  Median :0.0000   Median :652.0   Germany:2509                 Median :37.00
##  Mean   :0.2037   Mean   :650.5                                Mean   :38.92
##  3rd Qu.:0.0000   3rd Qu.:718.0                                3rd Qu.:44.00
##  Max.   :1.0000   Max.   :850.0                                Max.   :92.00
##      Tenure          Balance       NumOfProducts    HasCrCard
##  Min.   : 0.000   Min.   :     0   Min.   :1.00   Min.   :0.0000
##  1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00   1st Qu.:0.0000
##  Median : 5.000   Median : 97199   Median :1.00   Median :1.0000
##  Mean   : 5.013   Mean   : 76486   Mean   :1.53   Mean   :0.7055
##  3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:2.00   3rd Qu.:1.0000
##  Max.   :10.000   Max.   :250898   Max.   :4.00   Max.   :1.0000
##  IsActiveMember   EstimatedSalary
##  Min.   :0.0000   Min.   :    11.58
##  1st Qu.:0.0000   1st Qu.: 51002.11
##  Median :1.0000   Median :100193.91
##  Mean   :0.5151   Mean   :100090.24
##  3rd Qu.:1.0000   3rd Qu.:149388.25
##  Max.   :1.0000   Max.   :199992.48
```

Correlation

We can check is any of our predictor variables are highly correlated, we will remove any that are, as this will throw the accuracy of our model:

```
# analyze the bank influencing factors
cor_bank <- bank_churn_data_tbl %>%
  keep(is.numeric) %>%
  cor()

corrplot::corrplot(cor_bank, method = "circle")
```



All good here, no need to remove any of our numeric predictor variables.

Target Variable: Churn

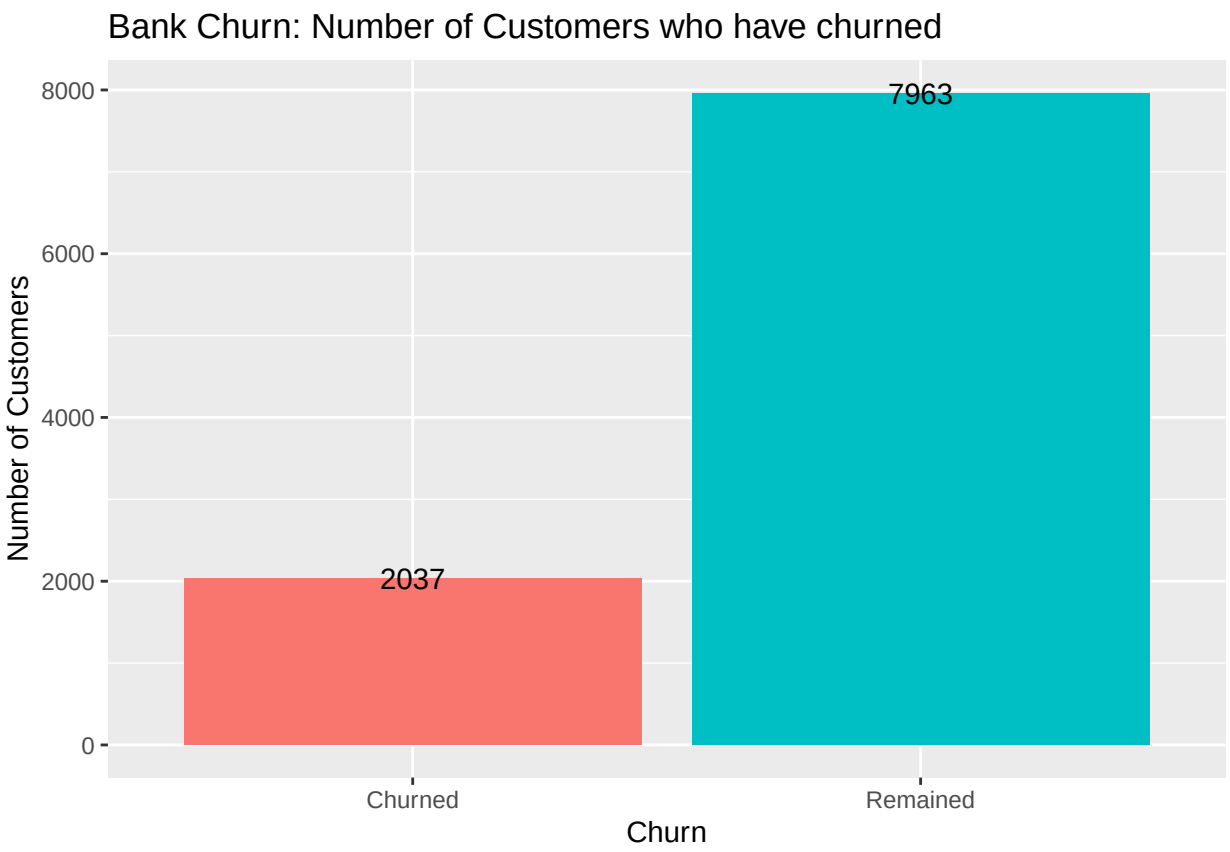
Churn: the target variable, I've changed the column name from 'Exited' to 'Churn,' as the analysis is about churn. A 1 in this cell tells us a customer has churned. 2037 of the total 10000 has churned.

```
bank_churn_n <- bank_churn_data_tbl %>%
  group_by(Churn) %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  tally()

bank_churn_n %>%
  ggplot(aes(Churn, n, fill = Churn)) +
  geom_col() +
  geom_text(label = bank_churn_n$n) +
  theme(legend.position = 'none') +
```

Churn	n
Churned	2037
Remained	7963

```
  labs(title = "Bank Churn: Number of Customers who have churned",
       y = "Number of Customers")
```




```
bank_churn_n %>%
  kableExtra::kable() %>%
  kableExtra::kable_styling()
```

Predictor variables

- **CreditScore**: ranges from 350 to 850, with a mean of 650.
- **Geography**: customers come from three countries from largest to smallest France, Germany and Spain.
- **Gender**: 5457 Men and 4543 Women
- **Age**: the customers age range from 18 to 92, the average is 38
- **Tenure**: how many year a bank customer range from 0 to 10, the average is 5.
- **Balance**: how much money does a customer have in the bank, range from Euro 0 to 250,898, the average is 76,486
- **NumOfProducts**: banking products per customer, range 1 to 4, average is 1.53
- **HasCrCard**: owns a credit card, 7055 customers of the 10000 total.

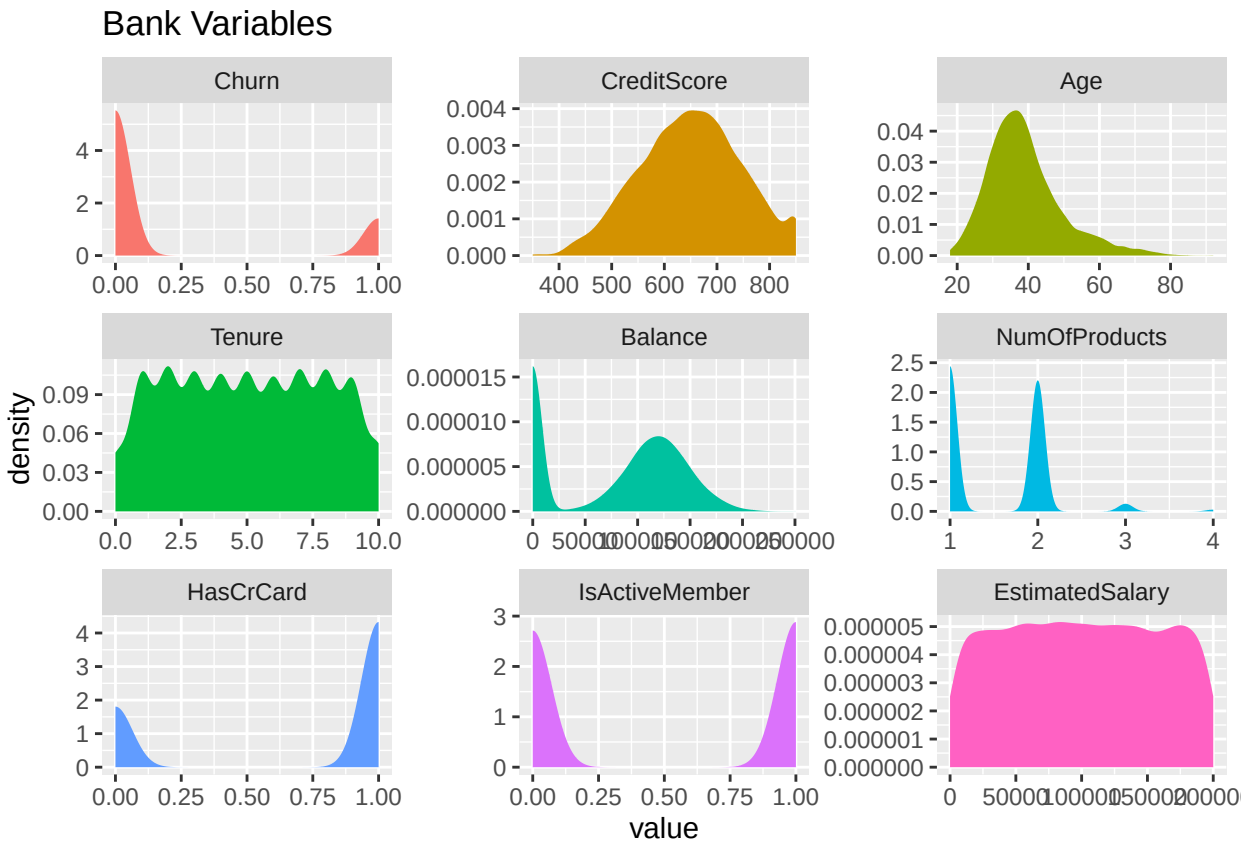
- **IsActiveMember**: presumably, makes regular deposits, 5151 of the 10000 total.
- **EstimatedSalary**: customers annual salary, range from Euro 11.58 to 199992, mean is 100090.

```
# lets melt the data so we can quickly display
melt.bank <- melt(bank_churn_data_tbl)
```

```
## Using Geography, Gender as id variables
```

```
#
p2 <- melt.bank %>%
  ggplot(aes(x = value, fill = variable)) +
  stat_density(show.legend = FALSE) +
  facet_wrap( ~variable, scales = "free") +
  labs(title = "Bank Variables")

p2
```



Categorical Predictor Variables

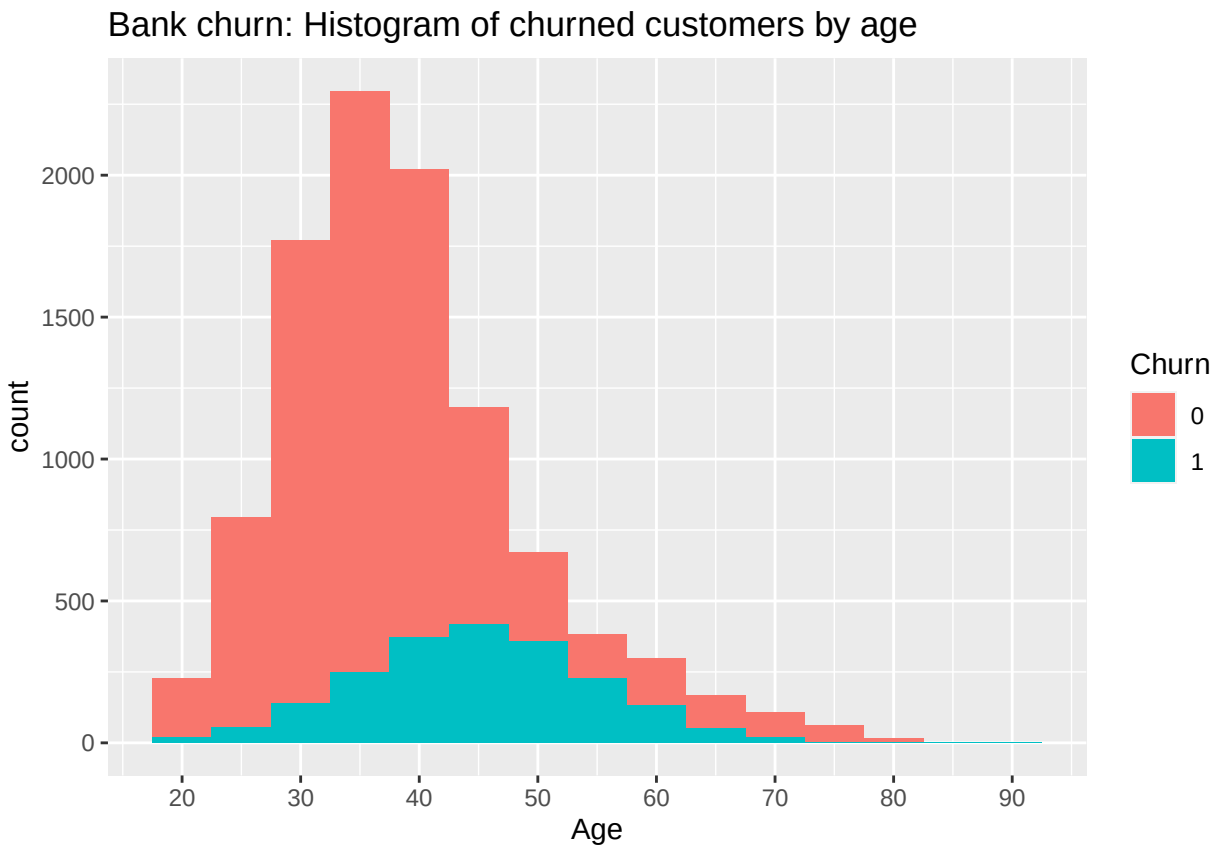
```
# 25 percent of females churned, a higher proportion of the males
bank_churn_data_tbl %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  dplyr::group_by(Gender, Churn) %>%
  tally() %>%
```

```
  ggplot(aes(Gender, n, fill = Churn )) +
  geom_bar(position="fill", stat="identity") +
  labs(title = "Bank churn: Ratio of churned customers by gender")
```

Gender Effect: women seem more likely to churn.

```
# 40 to 50 age range seems to be the age to churn, younger customers seem to be more stable. Could it b
age_hist_gg <- bank_churn_data_tbl %>%
  mutate(Churn = as.factor(Churn)) %>%
  ggplot(aes(x = Age, fill = Churn)) +
  geom_histogram(binwidth = 5, show.legend = TRUE) +
  scale_x_continuous(breaks = seq(0,100, by=10)) +
  labs(title = "Bank churn: Histogram of churned customers by age")

age_hist_gg
```

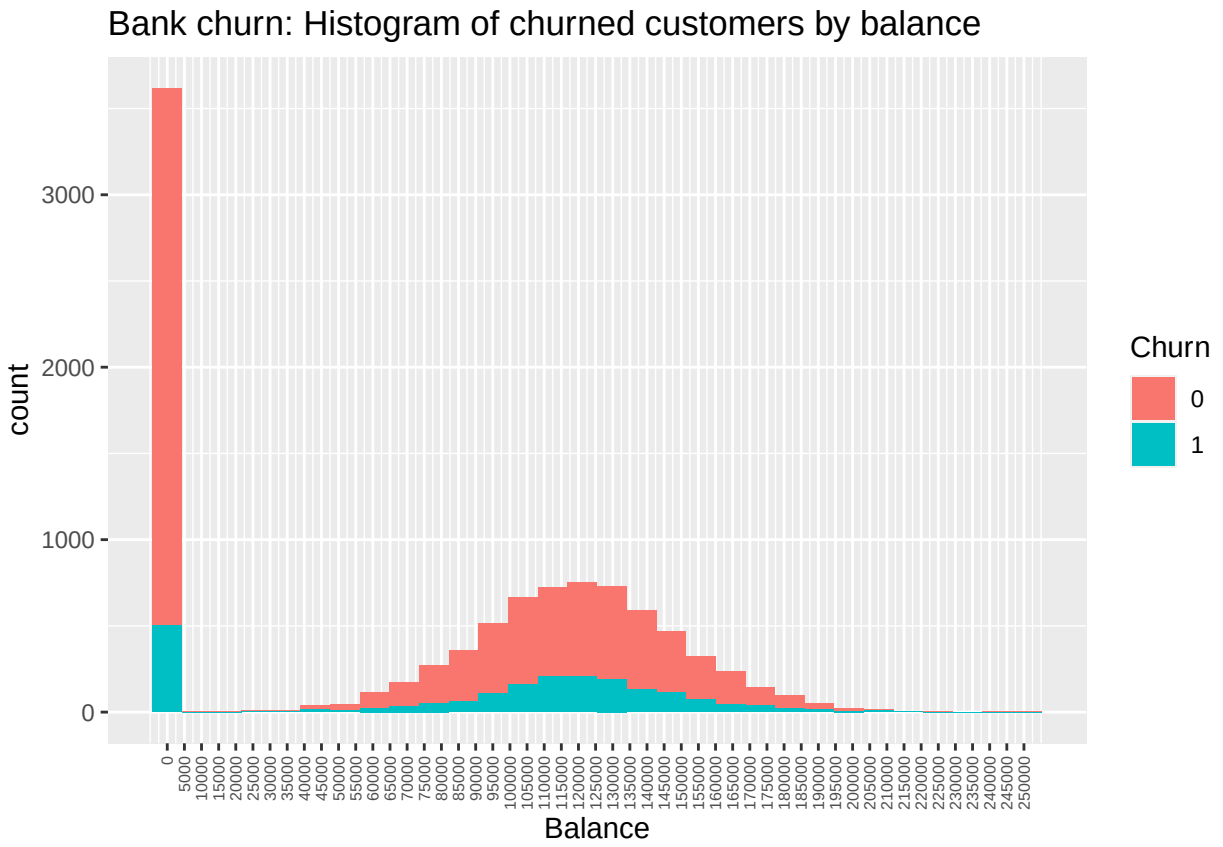


Age seems to be an significant: 40 to 50 age range seems to be the age to churn, younger customers seem to be more stable. Could it be older customers have done with saving and now need to invest their money?

```
bank_churn_data_tbl %>%
  mutate(Churn = as.factor(Churn)) %>%
  ggplot(aes(x = Balance, fill = Churn)) +
  geom_histogram(show.legend = TRUE) +
  scale_x_continuous(breaks = seq(0,max(bank_churn_data_tbl$Balance), by=5000)) +
```

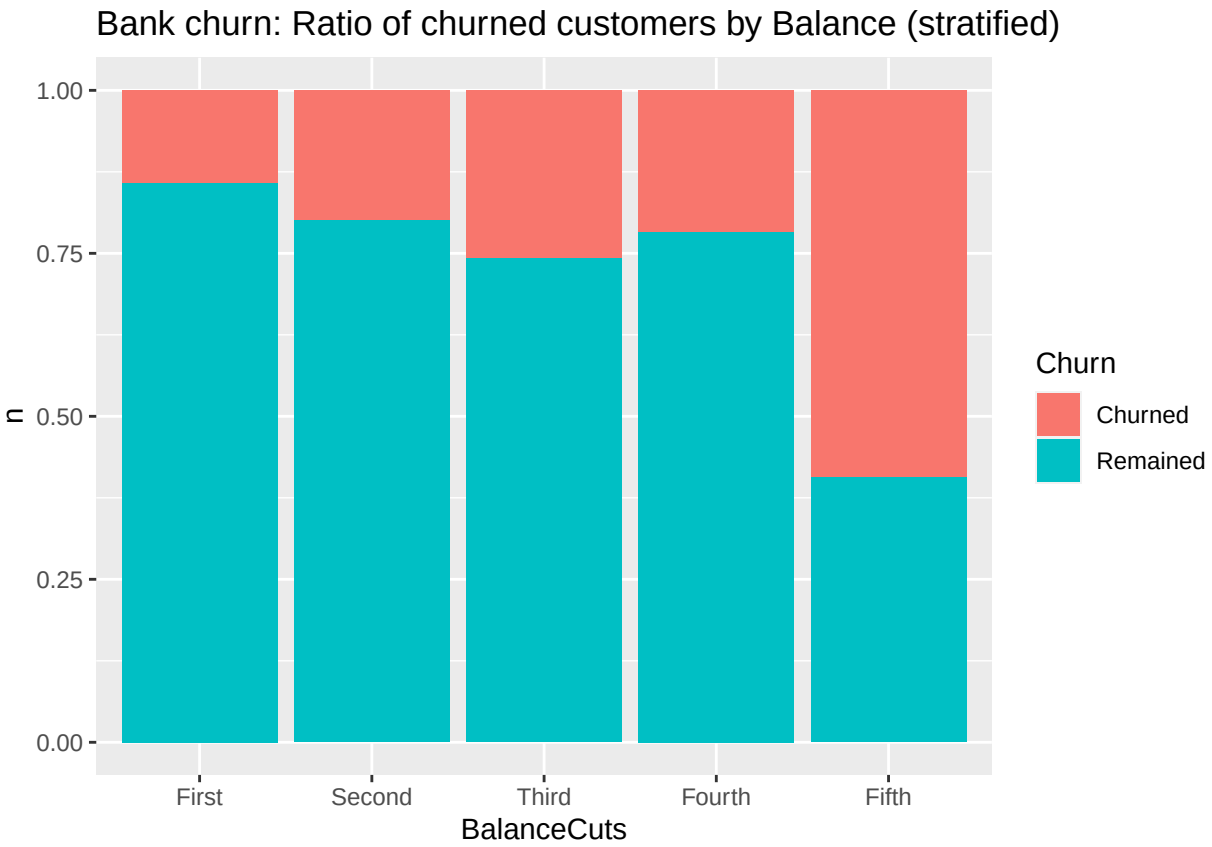
```
  labs(title = "Bank churn: Histogram of churned customers by balance") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 6))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



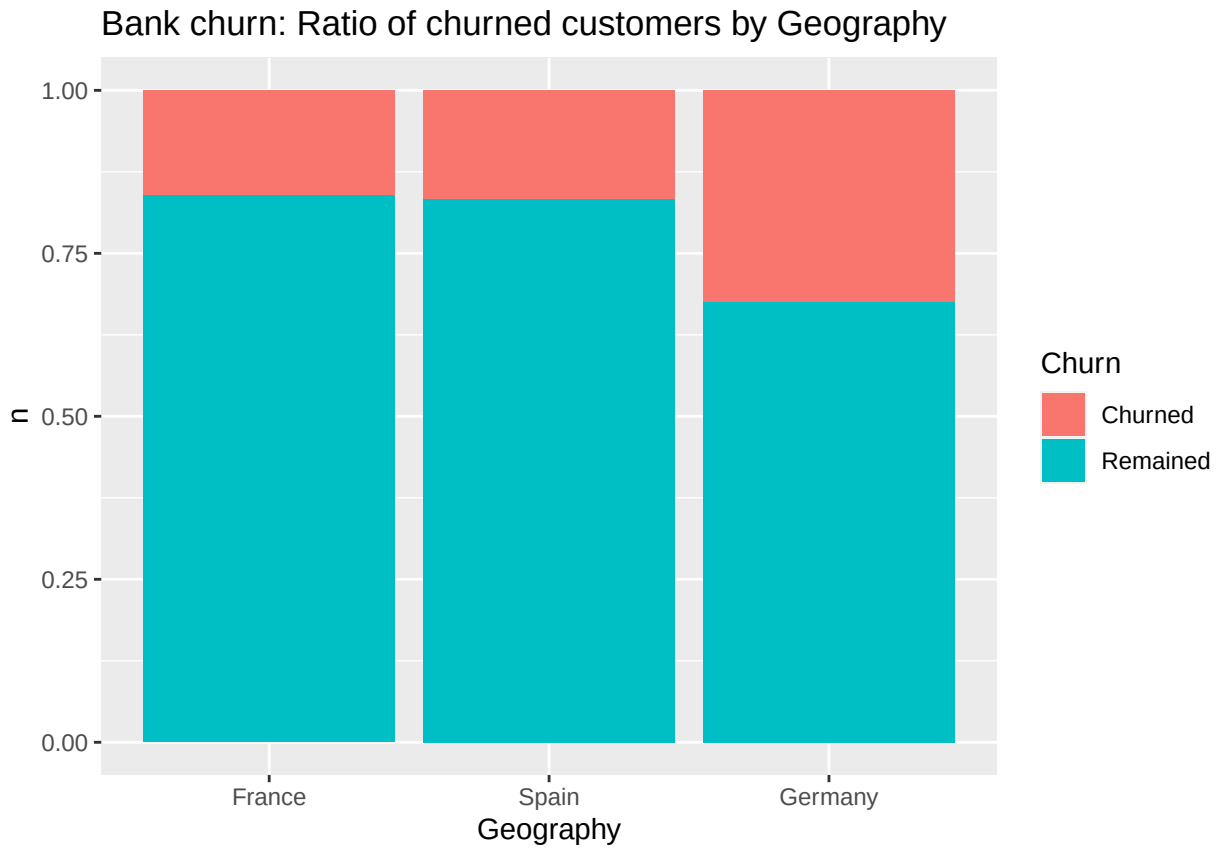
Balance is not massively significant I look's like churned by balance is only significant by 0 balance, it seems logical, customers are much more likely to churn, if they have no deposit. What happens if we bucket the balances?

```
# does the balance effect the churn?
bank_churn_data_tbl %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  mutate(BalanceCuts = cut(Balance, breaks = 5,
                           labels = c("First", "Second", "Third", "Fourth", "Fifth"))) %>%
  dplyr::group_by(BalanceCuts, Churn) %>%
  tally() %>%
  ggplot(aes(BalanceCuts, n, fill = Churn )) +
  geom_bar(position="fill", stat="identity") +
  labs(title = "Bank churn: Ratio of churned customers by Balance (stratified)")
```



Balance effect: If we stratify the Balance, the higher the balance, the more likely the customer is to churn.

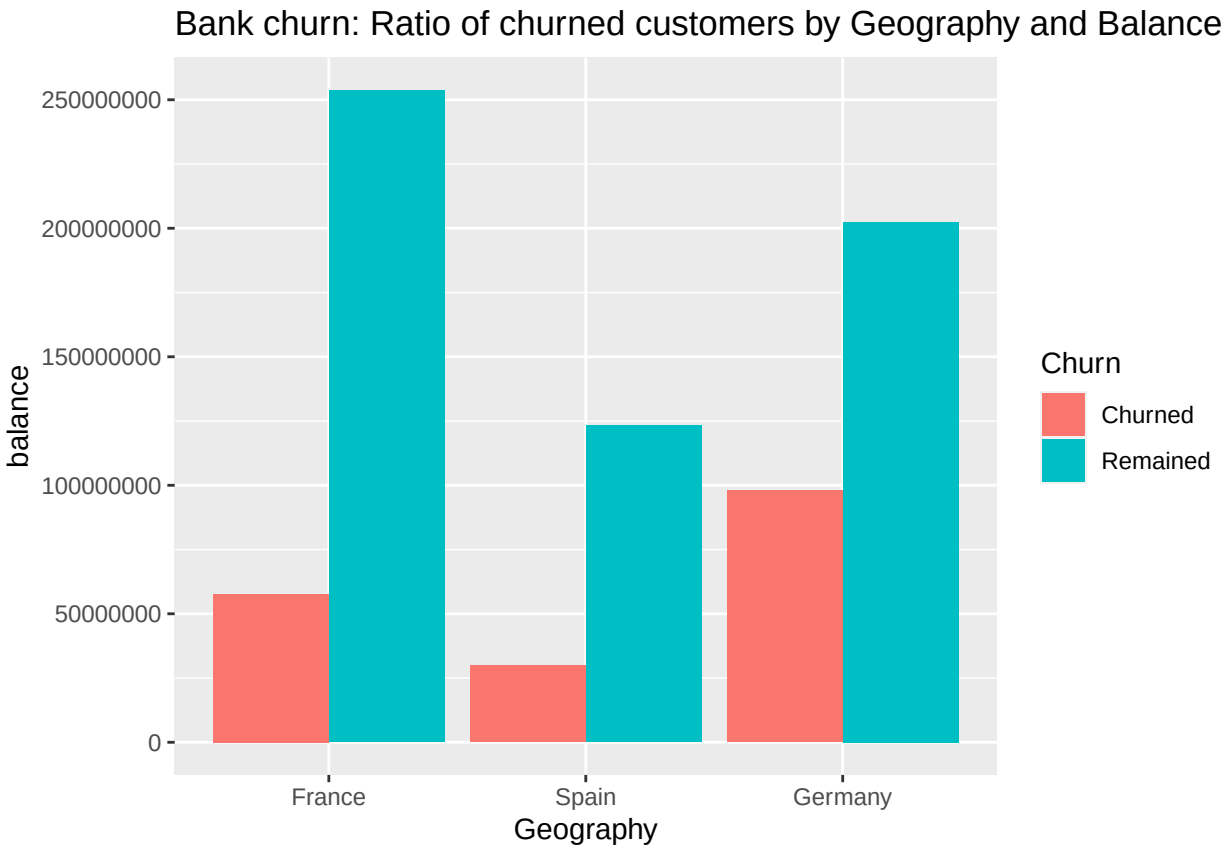
```
# Germany seems to be the churn capital
bank_churn_data_tbl %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  dplyr::group_by(Geography, Churn) %>%
  tally() %>%
  ggplot(aes(Geography, n, fill = Churn )) +
  geom_bar(position="fill", stat="identity") +
  labs(title = "Bank churn: Ratio of churned customers by Geography")
```



Geography seems significant: Germany seems to be the churn capital!

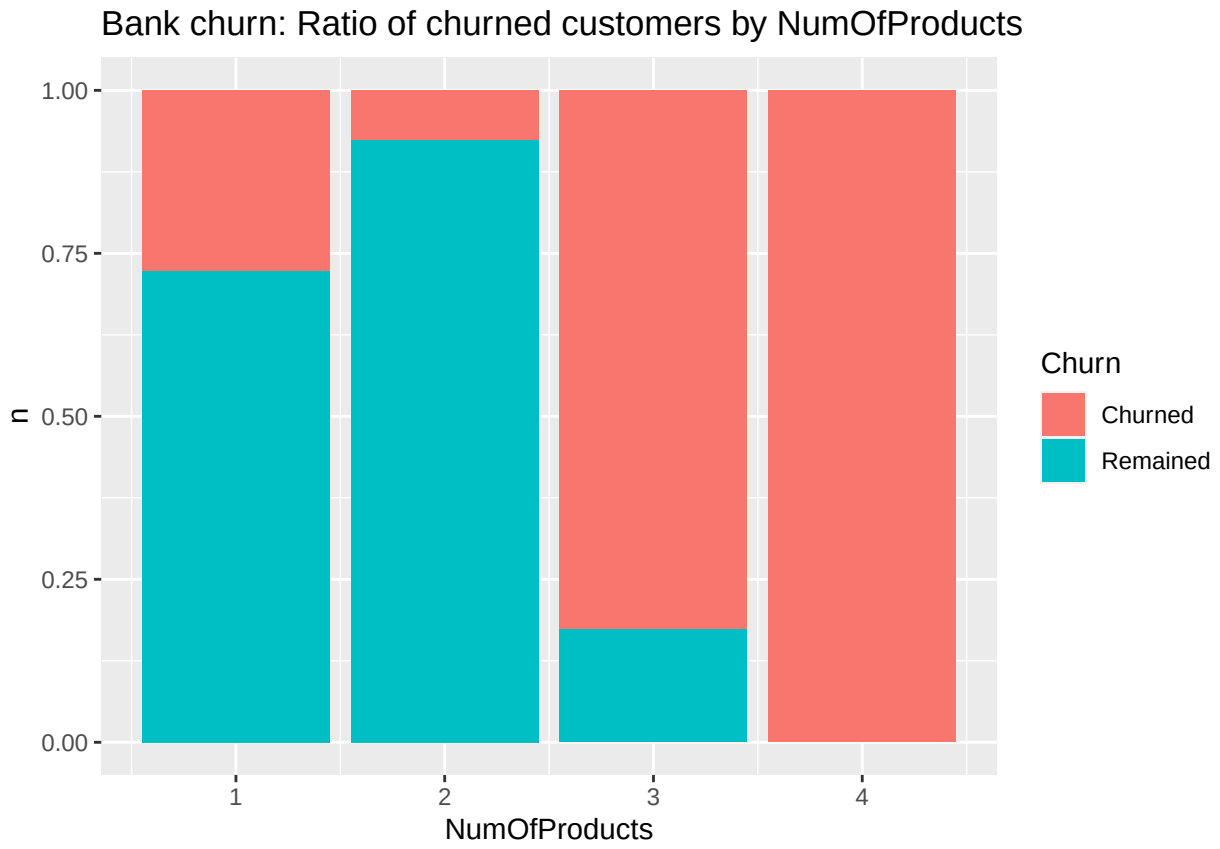
```
#
bank_churn_data_tbl %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  dplyr::group_by(Geography, Churn) %>%
  summarise(balance = sum(Balance)) %>%
  ggplot(aes(Geography, balance, fill = Churn )) +
  geom_bar(position="dodge", stat="identity") +
  labs(title = "Bank churn: Ratio of churned customers by Geography and Balance")
```

```
## `summarise()` has grouped output by 'Geography'. You can override using the `.groups` argument.
```



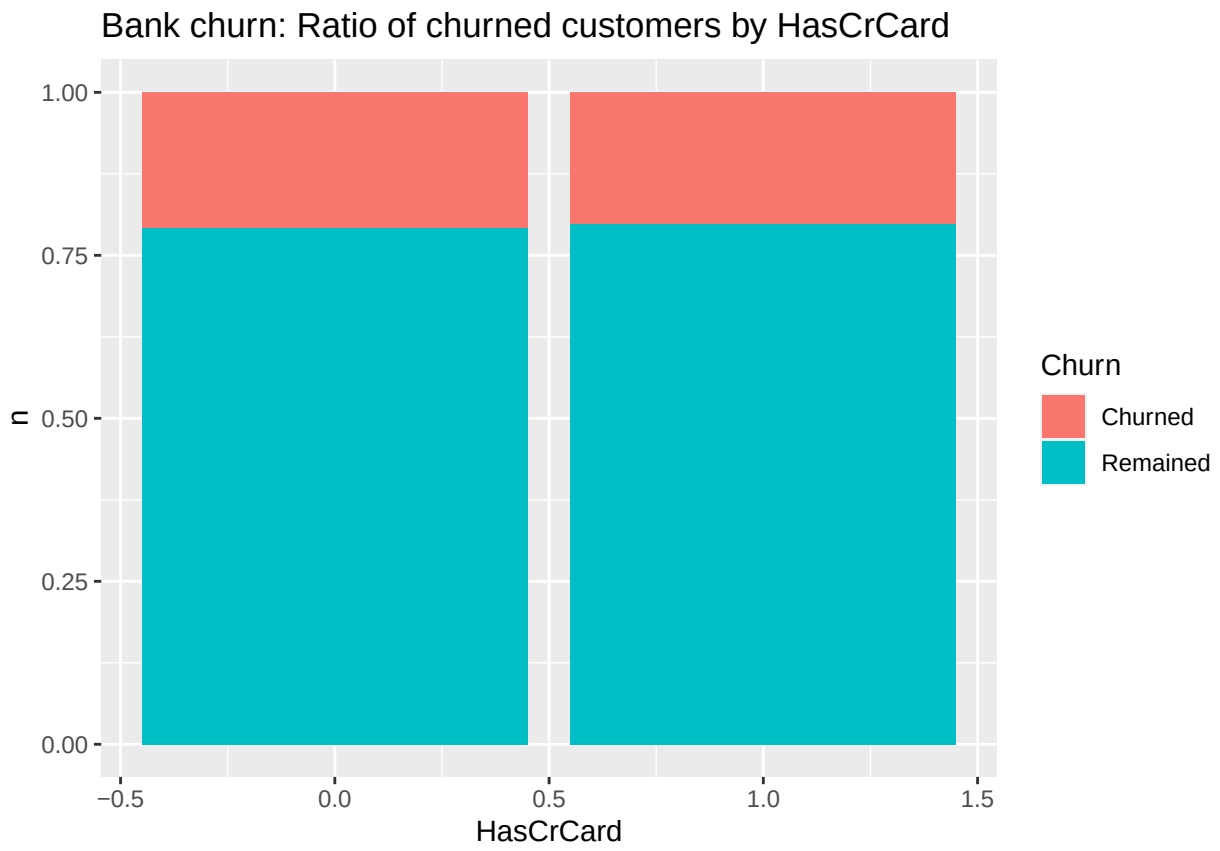
German balance: if we measure the balance, Germany has the biggest balance that churned.

```
# customers with more products churn more.
bank_churn_data_tbl %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  dplyr::group_by(NumOfProducts, Churn) %>%
  tally() %>%
  ggplot(aes(NumOfProducts, n, fill = Churn )) +
  geom_bar(position="fill", stat="identity") +
  labs(title = "Bank churn: Ratio of churned customers by NumOfProducts")
```



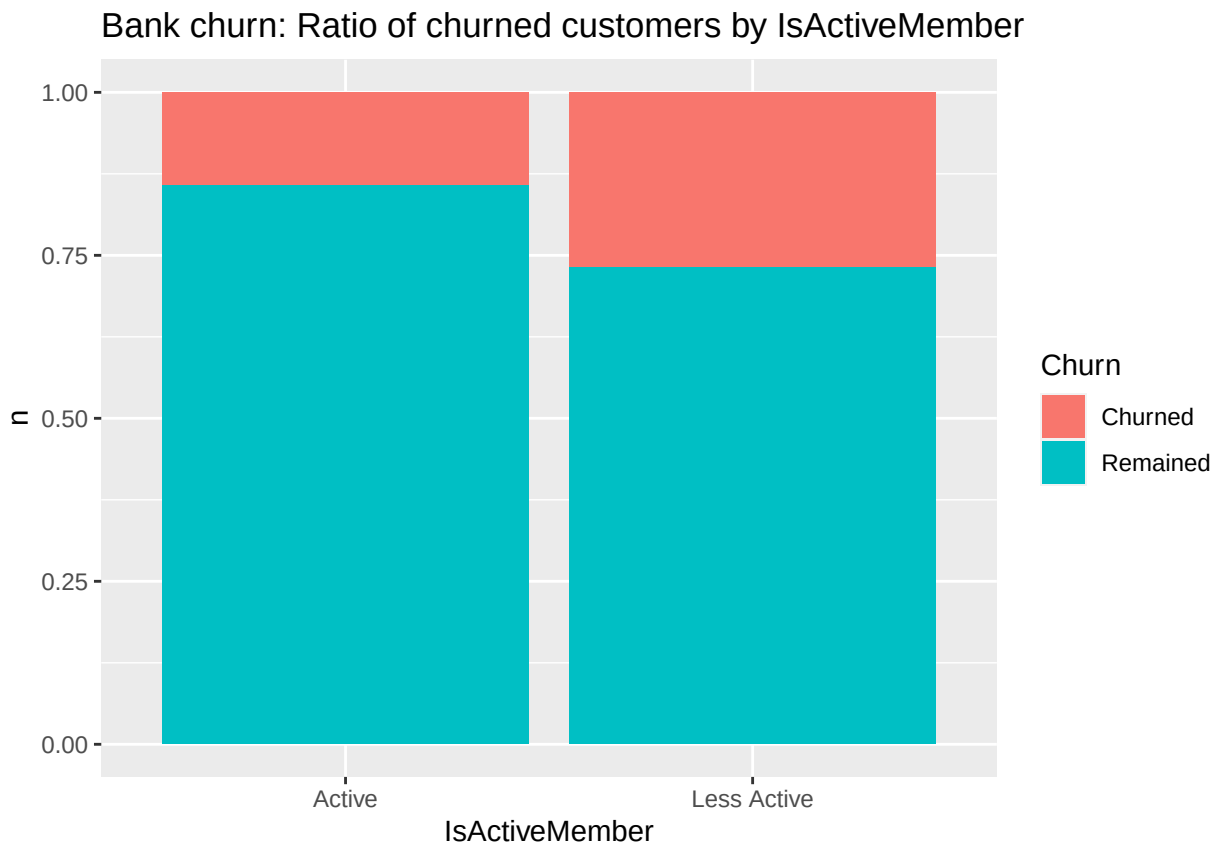
NumOfProducts effect: worryingly all customers who have 4 products have churned.

```
# credits card has no effect
bank_churn_data_tbl %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  dplyr::group_by(HasCrCard, Churn) %>%
  tally() %>%
  ggplot(aes(HasCrCard, n, fill = Churn )) +
  geom_bar(position="fill", stat="identity") +
  labs(title = "Bank churn: Ratio of churned customers by HasCrCard")
```



Credit Card: if a customer has a credit card or not, seems to have no/little effect.

```
# less active members seen to churn more
bank_churn_data_tbl %>%
  mutate(Churn = if_else(Churn == 1, "Churned", "Remained")) %>%
  mutate(IsActiveMember = if_else(IsActiveMember == 1, "Active", "Less Active")) %>%
  dplyr::group_by(IsActiveMember, Churn) %>%
  tally() %>%
  ggplot(aes(IsActiveMember, n, fill = Churn )) +
  geom_bar(position="fill", stat="identity") +
  labs(title = "Bank churn: Ratio of churned customers by IsActiveMember")
```



IsActiveMember effect: active customers with less activity seem more likely to churn.

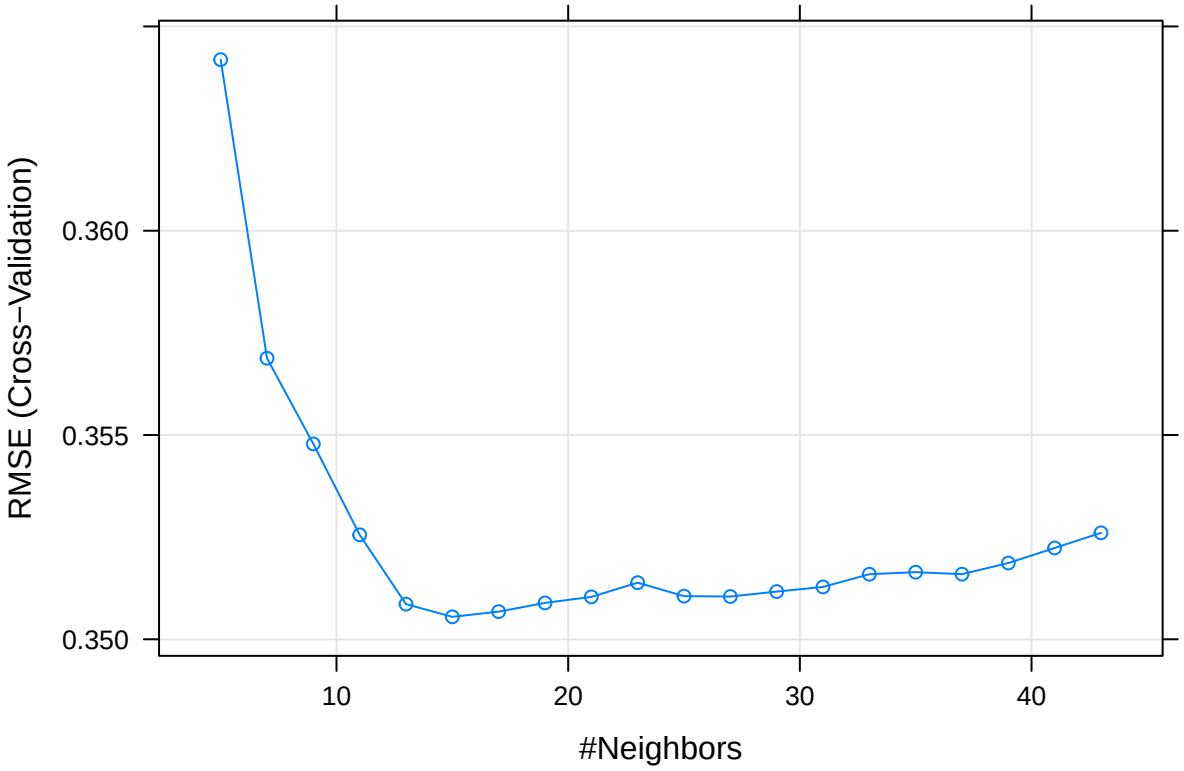
Modeling

We are targeting a numeric variable - Churn (0 or 1). Logistic regression is the chosen algorithm for such a categorization problem. I'd like the ability to choose the **cutoff point** for the model, as I will demonstrate, that financial cost of prediction may outweigh the accuracy of the model.

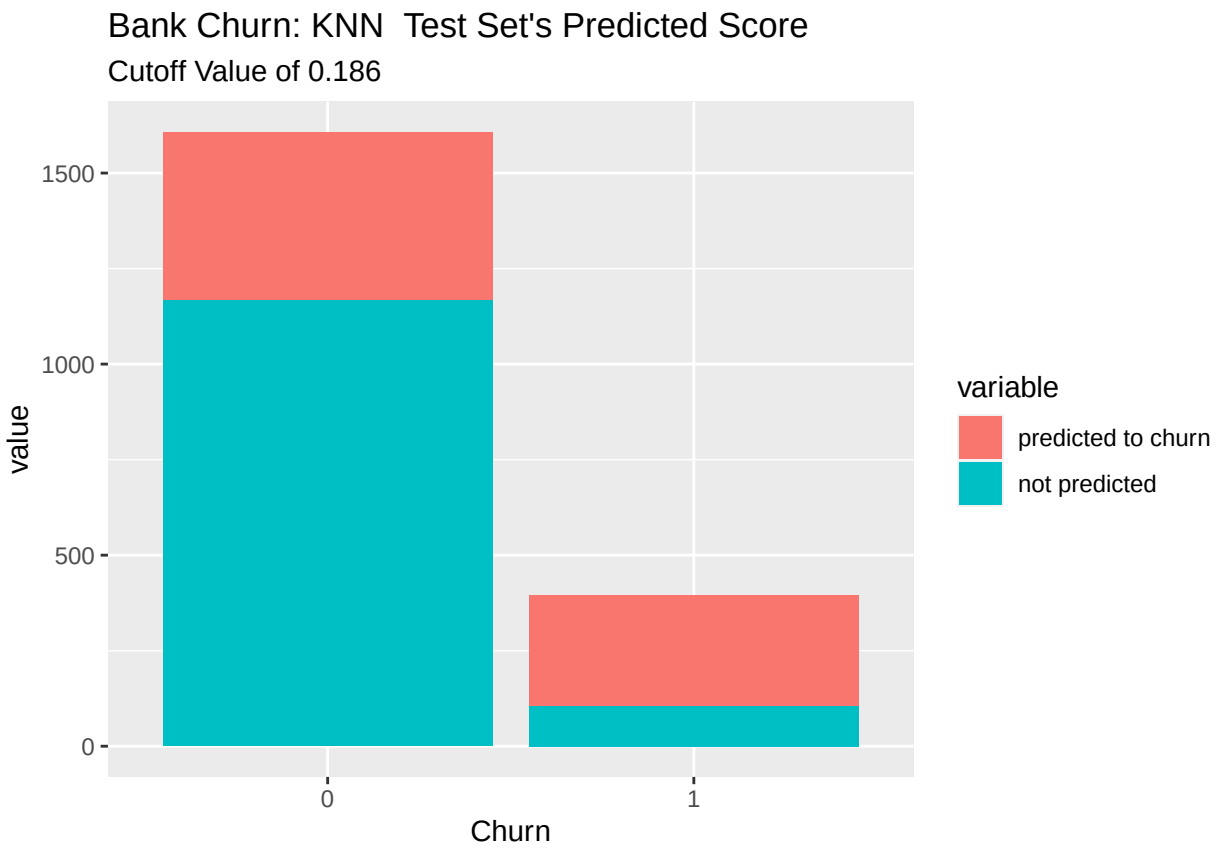
I will use a number of algorithms (as part of research, I evaluated more, but choose the four below based largely on accuracy and speed), to model Churn, these are:

- k-Nearest Neighbors algorithm (KNN) - Evelyn Fix and Joseph Hodges in 1951.
- Generalized Boosted Regression Models (GBM) - Friedman 1999.
- Multivariate adaptive regression spline (MARS) - Friedman in 1991
- General Logistic Model NET (GLM) - Freidman, Hastie and Tibshirani 2001.

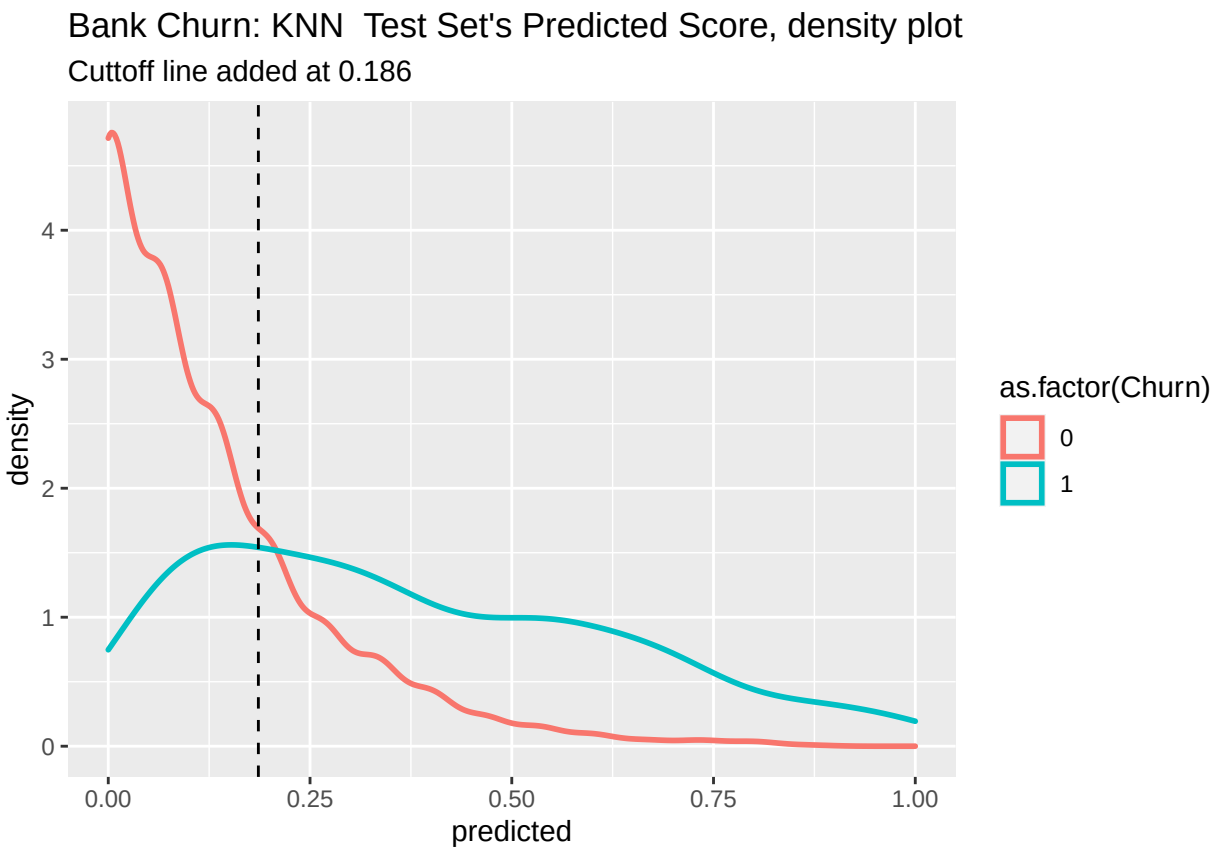
KNN




```
##    k
## 6 15
```



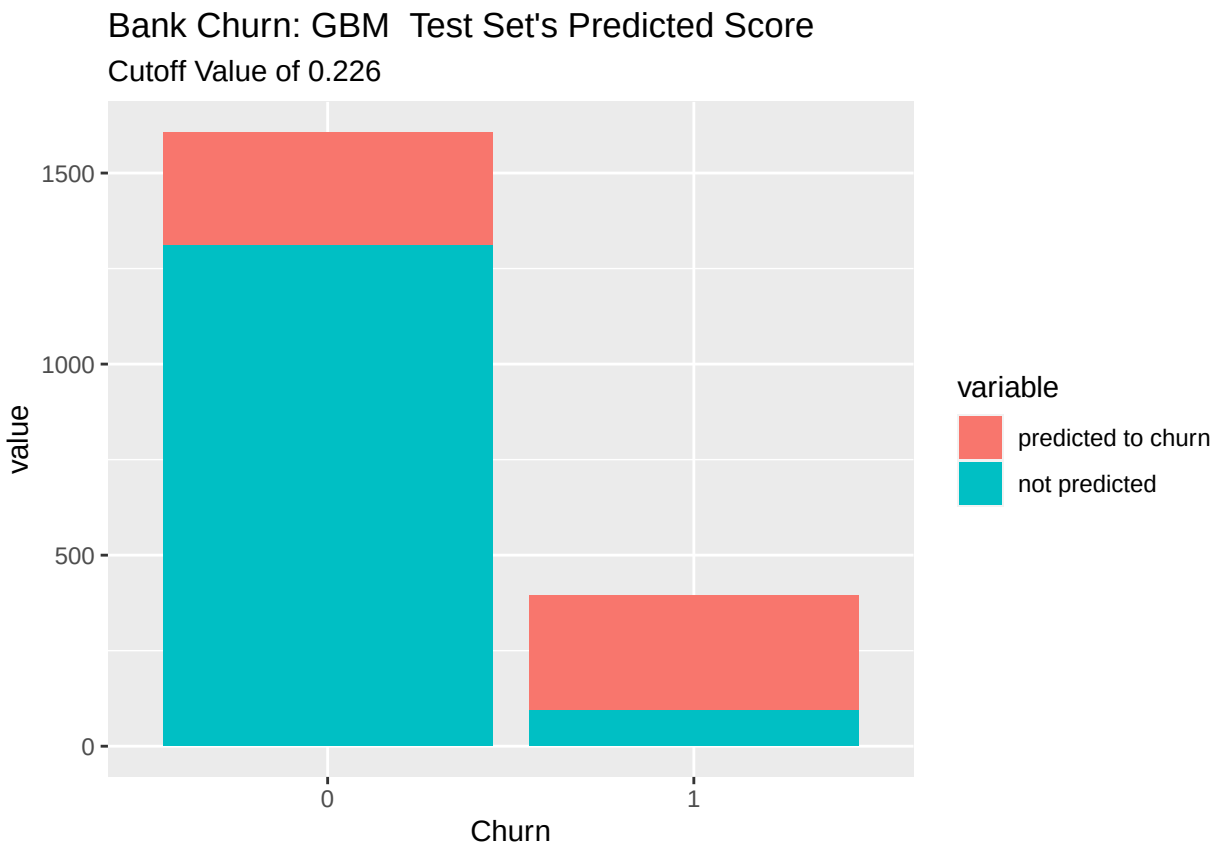
KNN predicted: we can see above how well KNN does with predicting churn (and mis-prediction).

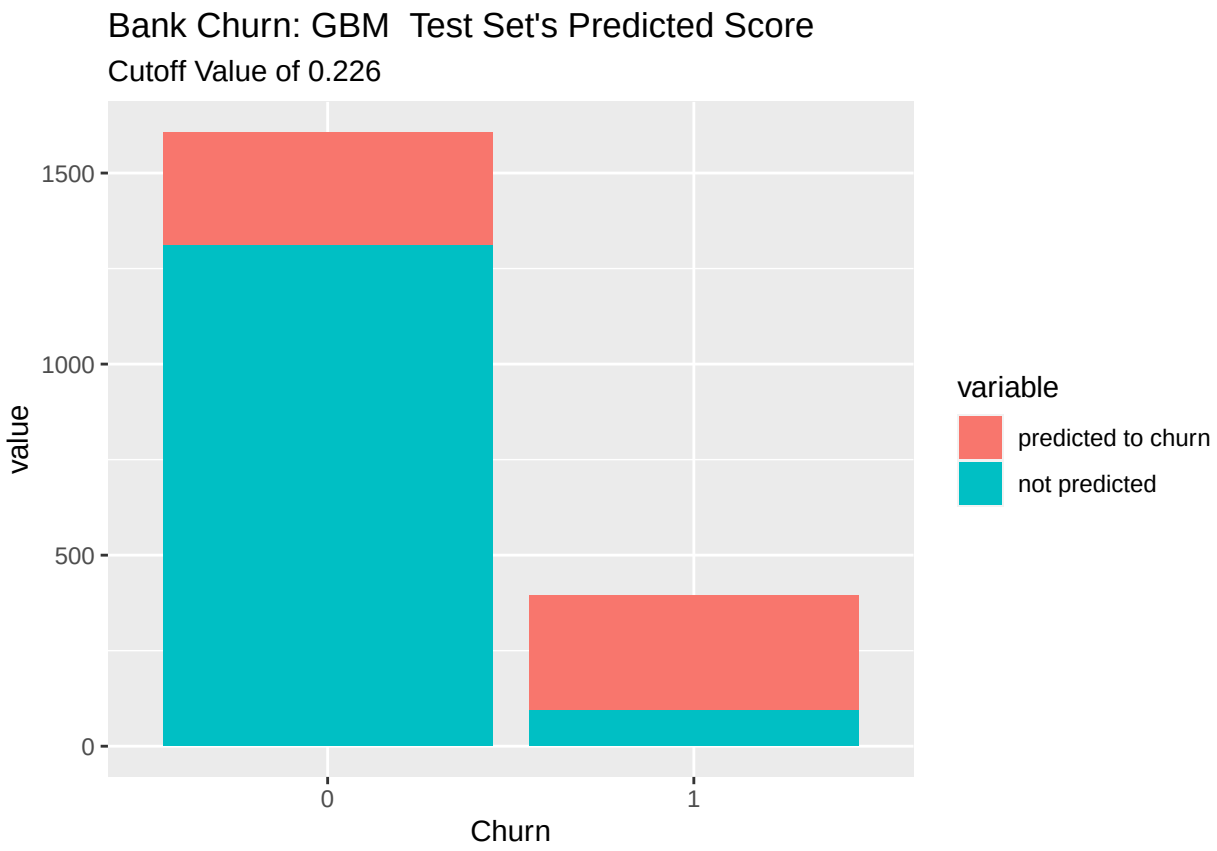


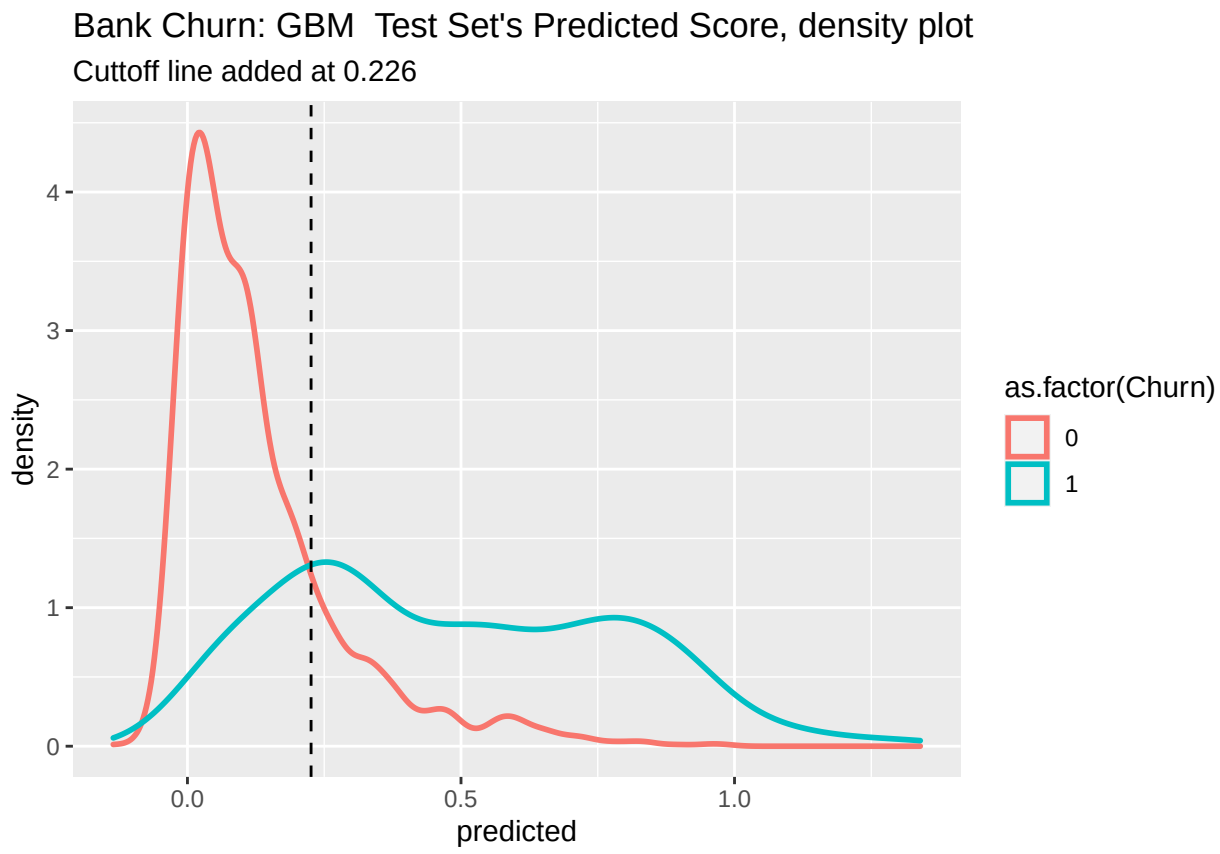
KNN density plot: For an ideal double density plot we want the distribution of scores to be separated, with the score of the negative instances to be on the left and the score of the positive instance to be on the right. In the KNN density plot, we can see that the negative and positives are not well distributed.

GBM

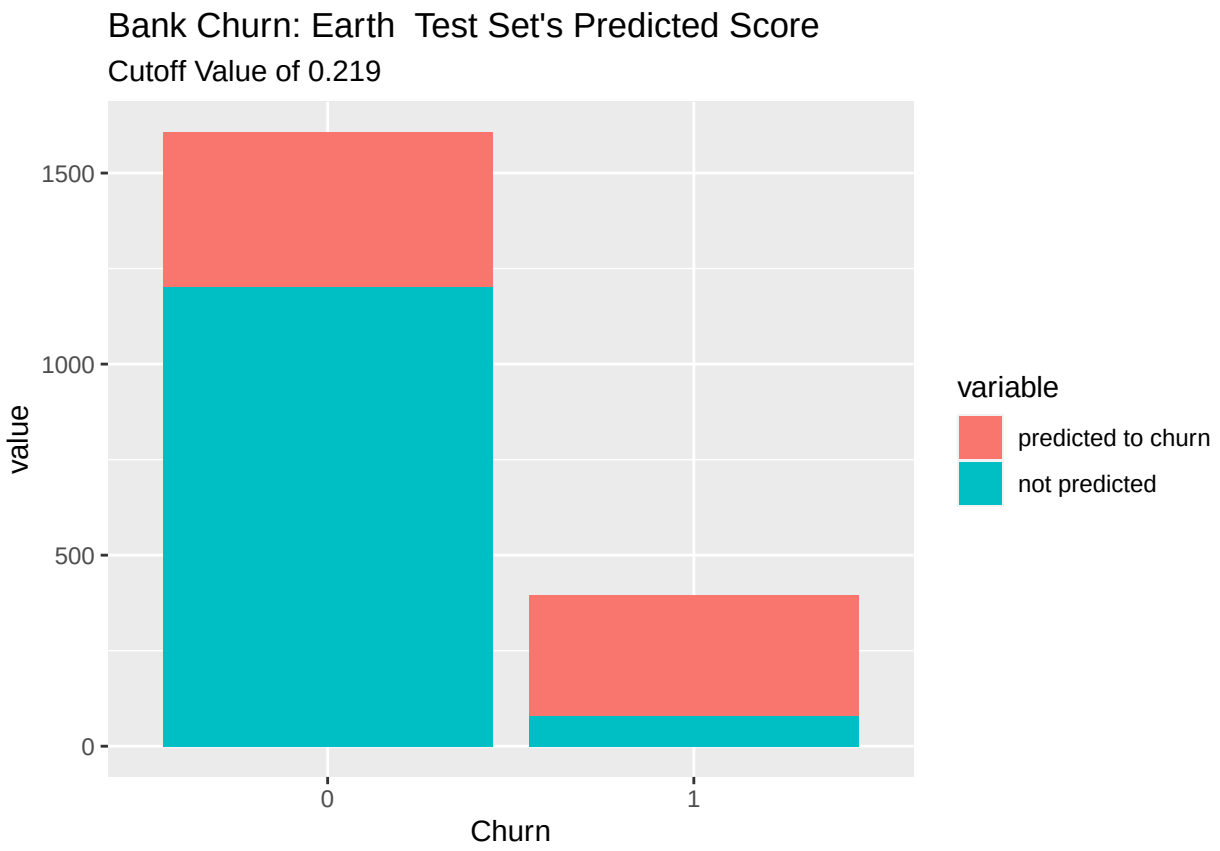
```
bank_churn_gbm
```


Bank Churn: GBM Test Set's Predicted Score
Cutoff Value of 0.226
1500
1000
500
0
value
0
1
Churn
variable
predicted to churn
not predicted

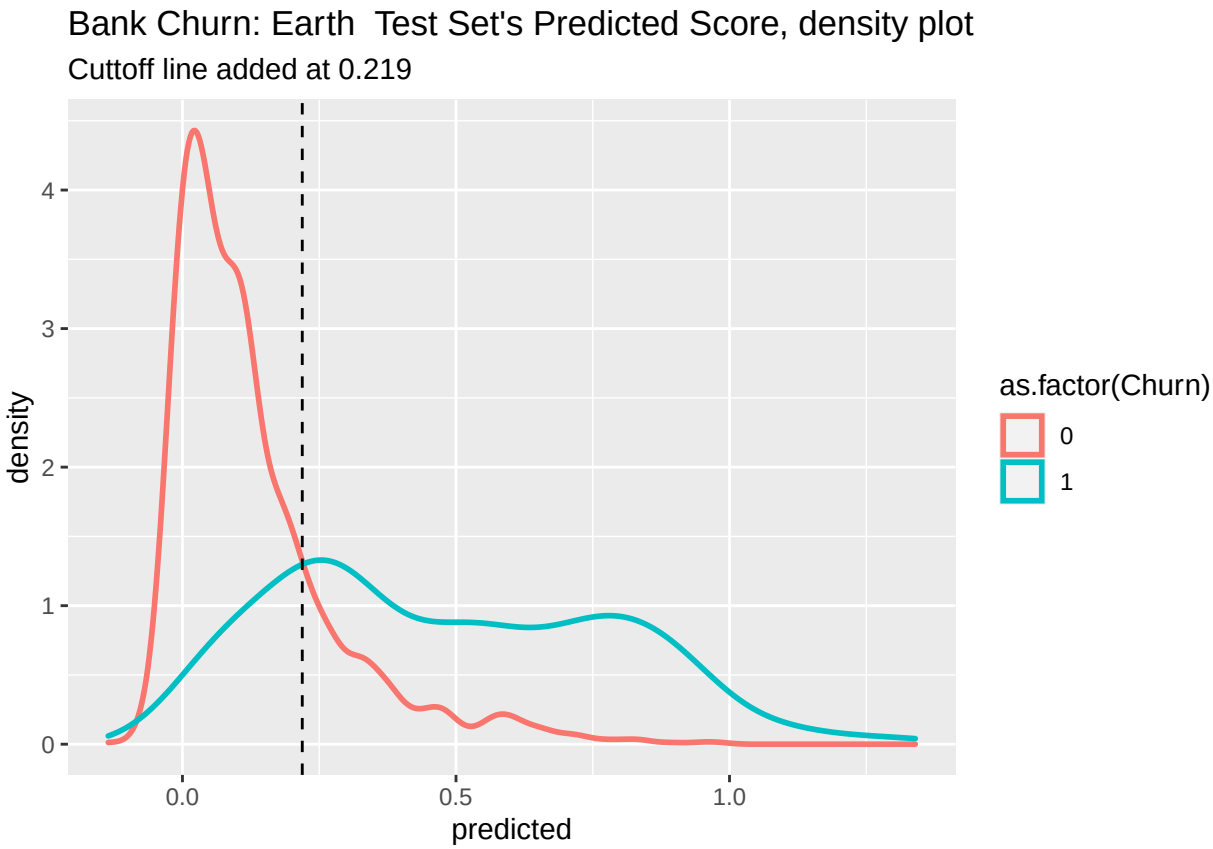

Bank Churn: GBM Test Set's Predicted Score
Cutoff Value of 0.226
1500
1000
500
0
value
0
1
Churn
variable
predicted to churn
not predicted

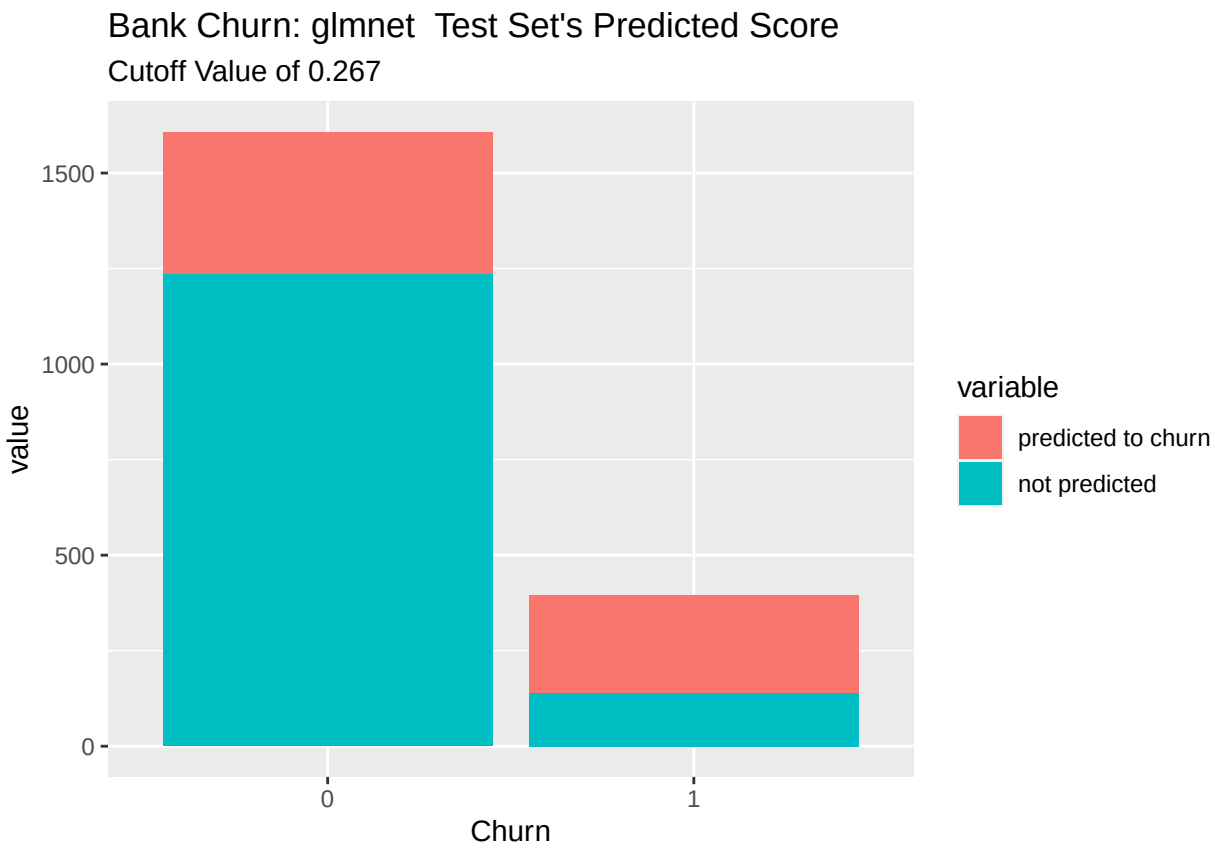


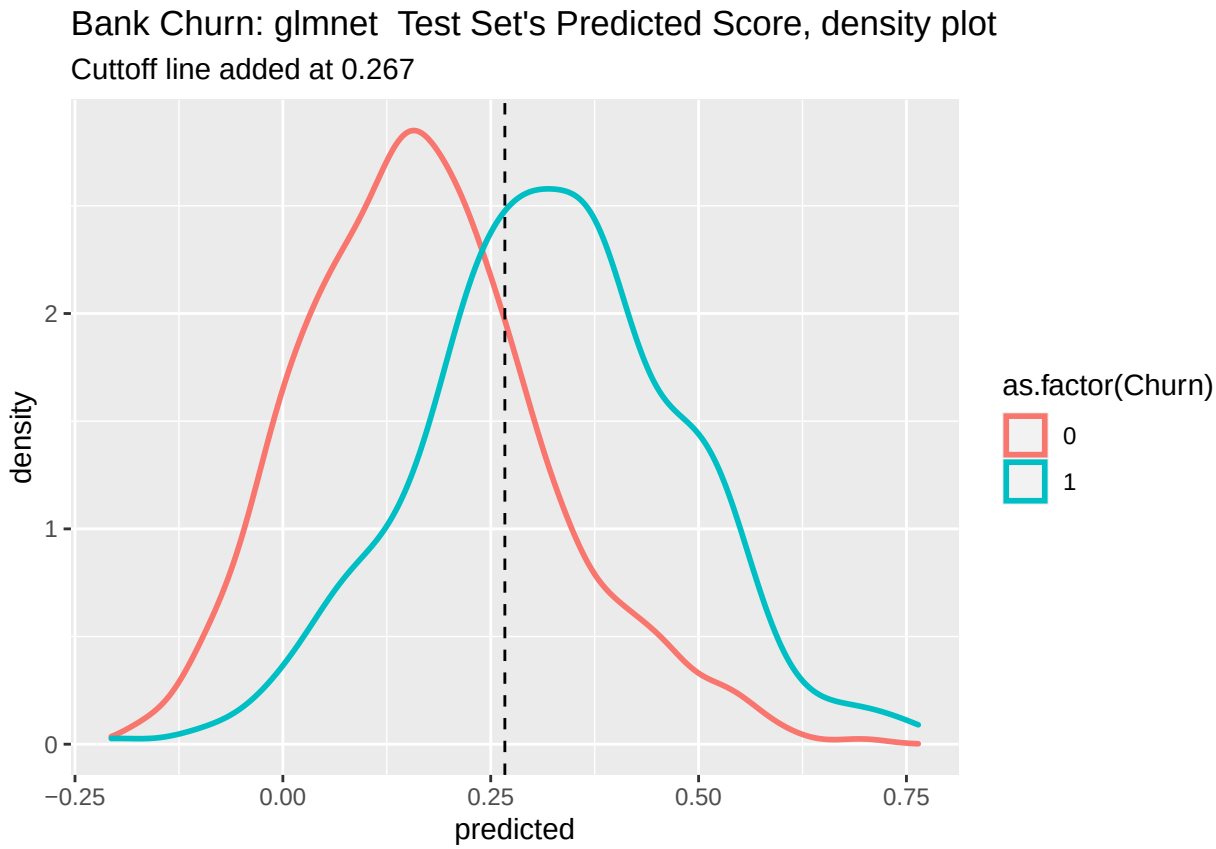

```
bank_churn_earth
```



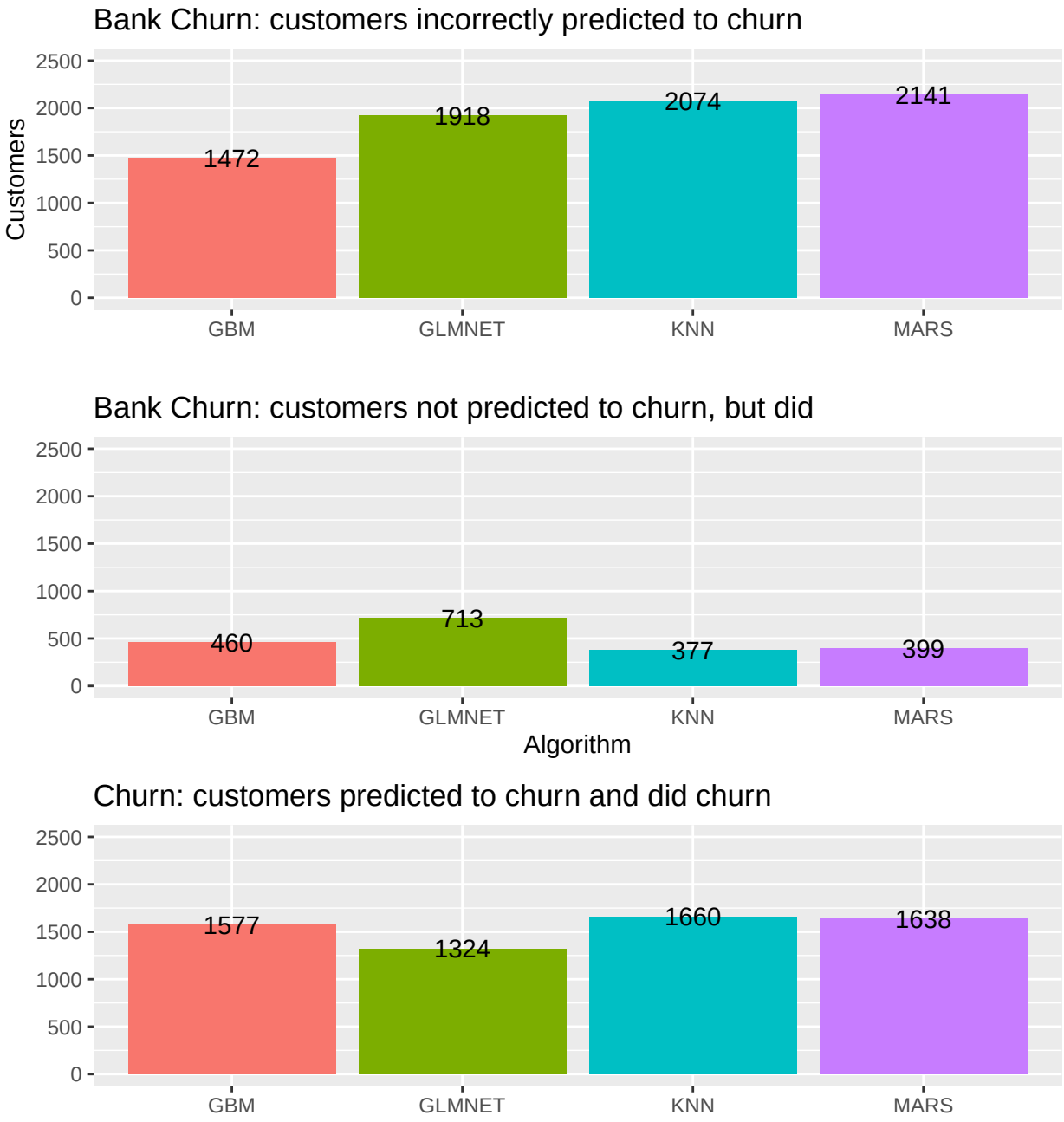

```
bank_churn_density_earth
```


Bank Churn: Earth Test Set's Predicted Score, density plot
Cuttoff line added at 0.219
density
predicted
0.0
0.5
1.0
0
1
2
3
4
as.factor(Churn)
0
1


Bank Churn: glmnet Test Set's Predicted Score
Cutoff Value of 0.267
1500
1000
500
0
value
0
1
Churn
variable
predicted to churn
not predicted


Bank Churn: glmnet Test Set's Predicted Score, density plot
Cuttoff line added at 0.267
density
predicted
0
1
2
−0.25
0.00
0.25
0.50
0.75
as.factor(Churn)
0
1

Model Accuracy



We can see that of our four algorithms, MARS predicted the most customers who would churn (1631), it also predicted a smaller amount of customers where predicted to churn but did not (400), however it predicted the largest amount of customers to churn, but who did not (2139).

Customer Lifetime Value

Customer Lifetime Value (CLV) is the present value of the future cash flows associated with a customer (Pfeifer et al, 2005). (Kahreh et al. 2014) propose a complex algorithm based on data we do not have. I've

Table 1: Bank Churn: customers incorrectly predicted to churn

type	customers
KNN	2074
GBM	1472
MARS	2141
GLMNET	1918

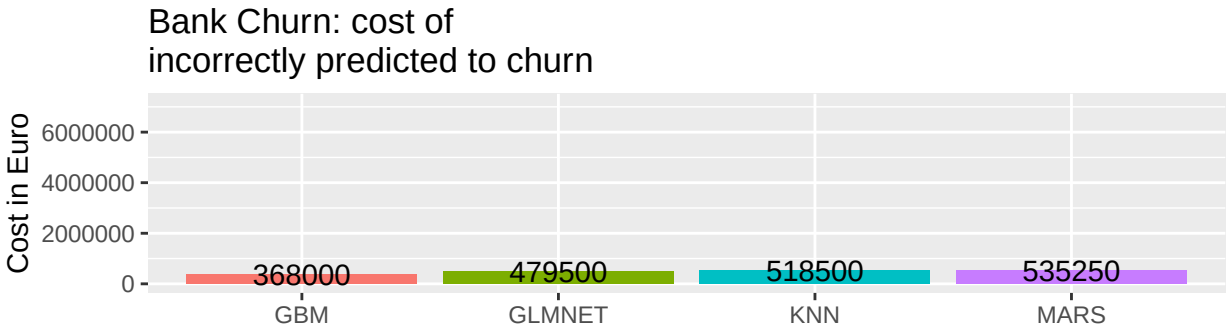
Table 2: Bank Churn: customers not predicted to churn, but did

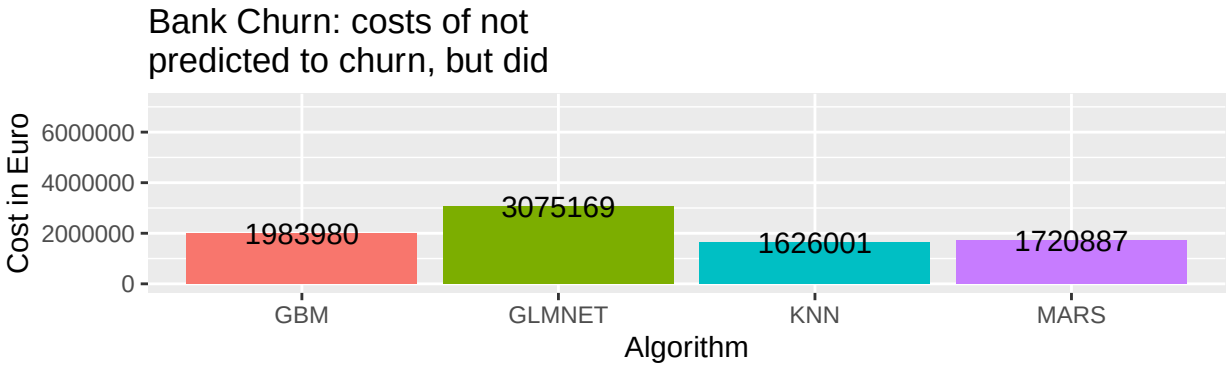
type	customers
KNN	377
GBM	460
MARS	399
GLMNET	713

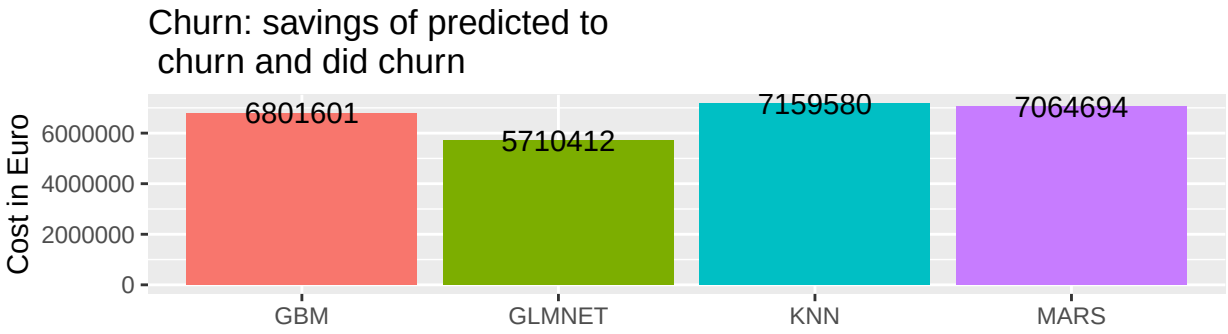
simplified their algorithm thusly:

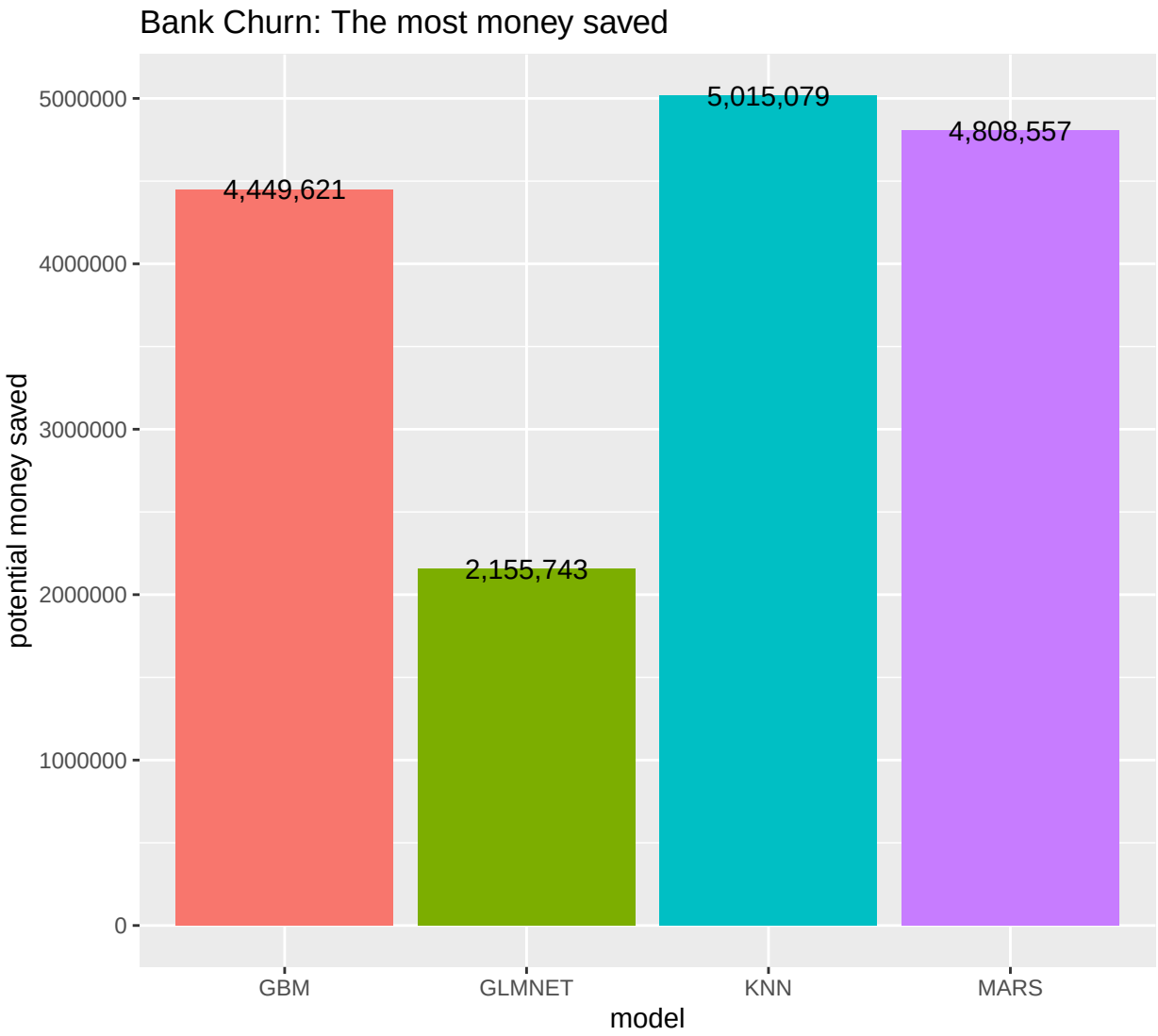
$$CustomerLifetimeValue = ((Balance * InterestRate) + (BankChargePerAnum * Tenure)) * CostOfKeepingCustomer$$

Now we will use our simple CLV to model the value of loosing our customers vs the predictive accuracy of our algorithm. We will assume that our bank makes a special offer of one year free banking (Euro 250) to keep our customers predicted to churn.


Bank Churn: cost of
incorrectly predicted to churn
Cost in Euro
6000000
4000000
2000000
0
368000
479500
518500
535250
GBM
GLMNET
KNN
MARS


Bank Churn: costs of not
predicted to churn, but did
Cost in Euro
6000000
4000000
2000000
0
1983980
3075169
1626001
1720887
GBM
GLMNET
KNN
MARS
Algorithm


Churn: savings of predicted to
churn and did churn
Cost in Euro
6000000
4000000
2000000
0
6801601
5710412
7159580
7064694
GBM
GLMNET
KNN
MARS



calculated as an average, that value was 4313. Cost of special offer was calculated as free banking for the year 250.

Summary: even though MARS is not the most effective, it could potentially save a bank most money.

Conclusion

I've been thinking about this cost of modeling for some time now, since I came across this problem in a professional context. Academically, we can use a confusion matrix, AUC or RMSE to calculate the value of a model using outcome as our only metric. But in a business context the cost is most important:

Potential Impact

This approach is not unique, many authors have grappled with this cost problem (Lemmens and Gupta 2020), (Hadden et al. 2007), (Stripling et al. 2018). Although the problem of explaining complex models is well studied (Canhoto and Clear 2020) state "managers are delaying the adoption of AI and ML because they are unsure how these technologies can help their firms."

Table 3: Comparing the Cost of Prediction

type	clv_saved	clv_lost	cost_of_keeping	potential_all_up_saving
KNN	7159580	1626001	518500	5015079
GBM	6801601	1983980	368000	4449621
MARS	7064694	1720887	535250	4808557
GLMNET	5710412	3075169	479500	2155743

I don't think we have much practical advise on explaining in simple terms, to busy stakeholders, the value of machine learning to a business. Here I think I can offer some advise:

- talk money to business people
- keep your presentation short (5 slides) *you can go deeper to follow up on any questions that arise*
- keep the messaging tight, *AUC* don't rock, for a business audience, don't talk density plots, talk money.

Limitations

The cost model I used is limited: I assumed CLV and Cost of Keeping a Customer, in real world context, we would have more variables to model against. I also modeled an average CLV, instead calculating individual value of each customer.

Future Work

I came across one **R** library that deals with cost of cost sensitive classification (Lang et al. 2019), I'd like to apply that knowledge to this problem. I'd also like to submit a paper on this work.

Thank you for your time in reading this work.

References

Canhoto, Ana Isabel, and Fintan Clear. 2020. "Artificial Intelligence and Machine Learning as Business Tools: A Framework for Diagnosing Value Destruction Potential." *Business Horizons* 63 (2): 183–93.

Frank Sherlock, Callminer. 2018. "Avoidable Customer Churn Costs British Businesses £25 Billion." https://www.realwire.com/releases/Avoidable-customer-churn-costs-British-businesses-25-billion.

Hadden, John, Ashutosh Tiwari, Rajkumar Roy, and Dymitr Ruta. 2007. "Computer Assisted Customer Churn Management: State-of-the-Art and Future Trends." *Computers & Operations Research* 34 (10): 2902–17.

Kahreh, Mohammad Safari, Mohammad Tive, Asghar Babania, and Mostafa Hesan. 2014. "Analyzing the Applications of Customer Lifetime Value (CLV) Based on Benefit Segmentation for the Banking Sector." *Procedia-Social and Behavioral Sciences* 109: 590–94.

Lang, Michel, Martin Binder, Jakob Richter, Patrick Schratz, Florian Pfisterer, Stefan Coors, Quay Au, Giuseppe Casalicchio, Lars Kotthoff, and Bernd Bischl. 2019. "mlr3: A Modern Object-Oriented Machine Learning Framework in R." *Journal of Open Source Software*, December. https://doi.org/10.21105/joss.01903.

Lemmens, Aurélie, and Sunil Gupta. 2020. "Managing Churn to Maximize Profits." *Marketing Science* 39 (5): 956–73.

Stripling, Eugen, Seppe vanden Broucke, Katrien Antonio, Bart Baesens, and Monique Snoeck. 2018. "Profit Maximizing Logistic Model for Customer Churn Prediction Using Genetic Algorithms." *Swarm and Evolutionary Computation* 40: 116–30.