**MLOps & Big Data – Final Project**

**Defining the Problem Statement**

The project revolves around a dataset that belongs to a European Bank and their customers. The bank has been seeing high attrition rates in recent times and want to take some actions to rectify the situation. We have the following 3 datasets available to us,

* Churn Dataset – Contains demographic information about 10,000 customers from the bank and their churn status. This is a sample of the bank’s total consumer base and indicative of the target population
* Churn Customers Outreach Dataset – Contains information about the income and spending of ~200 customers who stopped using the bank’s services. The bank would like to understand the reasons that motivated the exits but only wants to reach out to the subset of people who would be the most profitable to retain in the future.
* Ad Campaign Dataset – Once the bank identified the reasons for churn, they started an Ad Campaign with 10 different ads being shown to users. They would like to understand which one of the ads has the best CTR so that they can eliminate poorly performing ads and only focus on the ones doing well.

Through this project, we’ll be tackling the following problem statements,

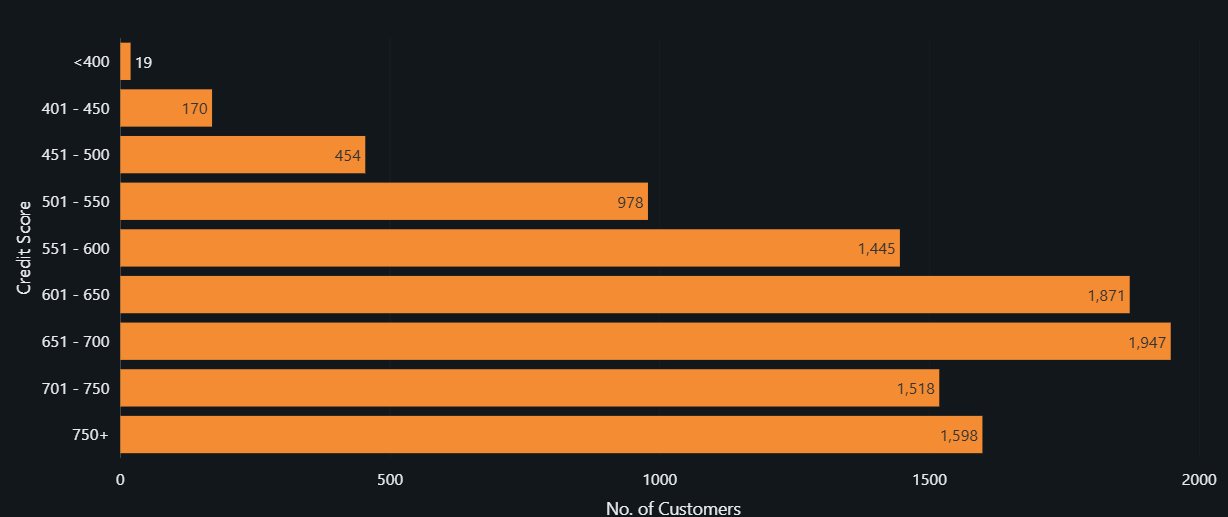
1. Understanding the bank’s ideal customer demographic based on the churn dataset. This will help them identify what kind of people are likely to use their services and therefore target their acquisition campaigns
2. Defining Churn Predictions through demographics and account metadata. This can act as an early indicator for predicting which customers are likely to churn in the future, allowing the bank to step in and avoid that from happening
3. Identify the most profitable cluster of churned customers for qualitative interviews to understand why High Value Customers stopped using the bank’s services.
4. Perform an analysis to identify the best performing Ad from the Ad Campaign.

**Identifying the Bank’s Customer Demographic**

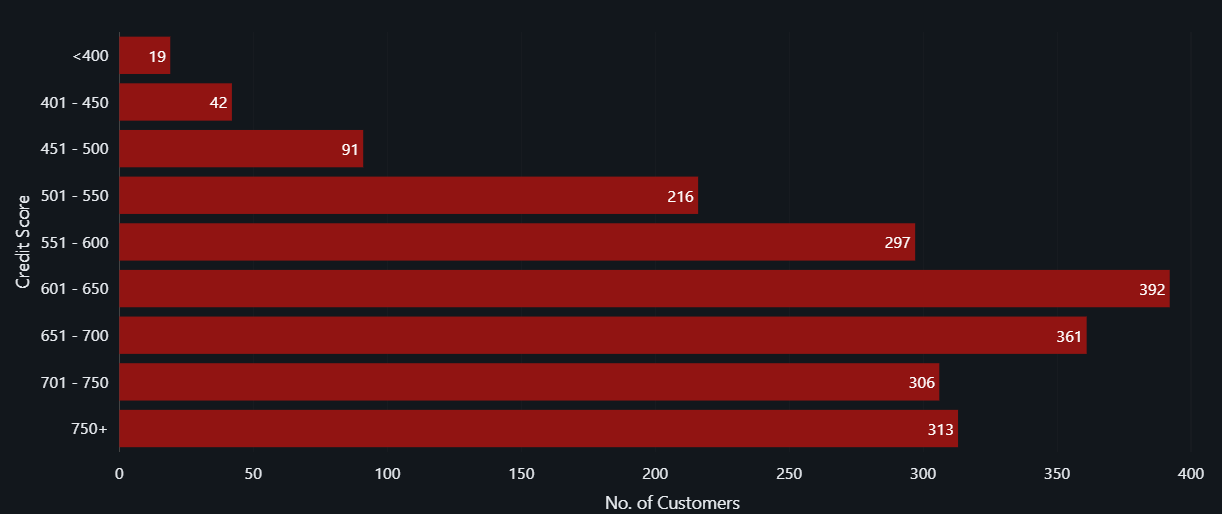
The very first thing we wish to address in this project, is understanding what a typical customer for our bank looks like. We have some demographic information related to each of the 10,000 customers such as age, gender, country etc along with some bank related metadata such as account balance, credit score, credit card ownership etc.

Based on these fields and the way that the customers are distributed across them, we can come up with the bank’s most common demographic. Checking this across 2 cohorts of customers – all 10,000 customers v/s customers who have exited the bank’s services will allow us to see if there are some clear distinctions allowing us to identify churn pre-emptively.

Let’s start by looking at the credit score distribution for both the cohorts,



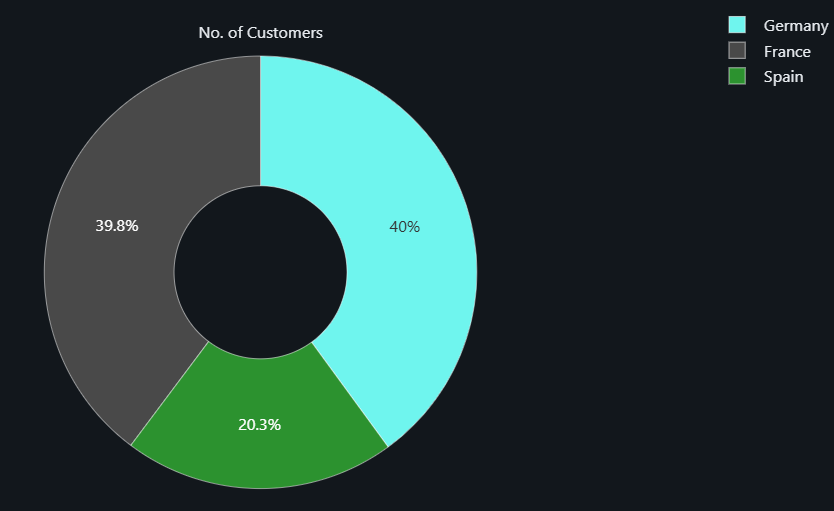
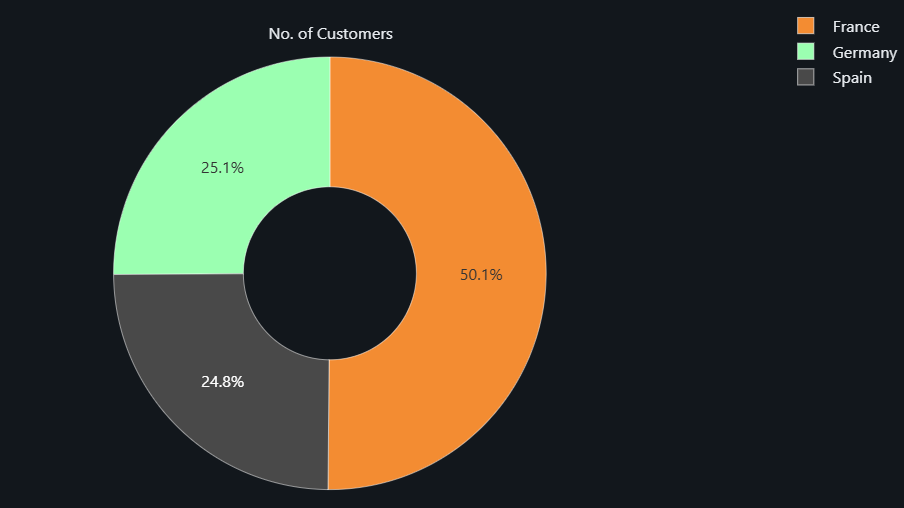
***Credit Score Distribution for ALL Customer***



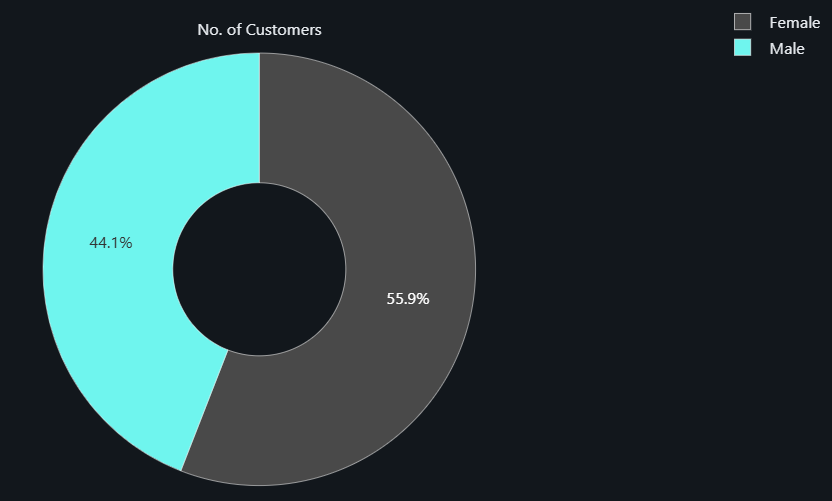
***Credit Score Distribution for CHURNED Customer***

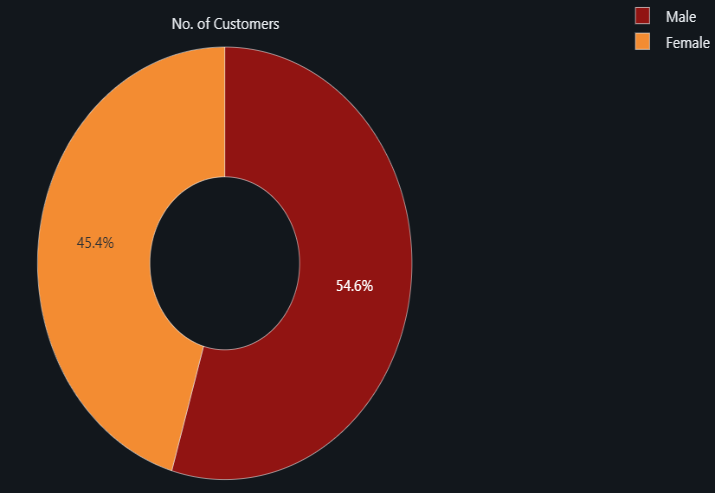
The one clear observation that can be made here is that all 19 users with a Credit Score < 400 ended up churning out. This means that users with a credit score lower than 400 are at a higher likelihood of churning out.

Similarly, we can also look at the distribution across age, country, tenure with the bank, number of banking products/services being used etc to find out other clear patterns.

Looking at the distribution across countries for both cohorts – All Customers (Left) and Churned Customers (Right)

We can see that there is a difference in the distribution of the customers. This is due to a higher proportion of German customers churning out. This means that German Customers are more likely to churn than French or Spanish customers.

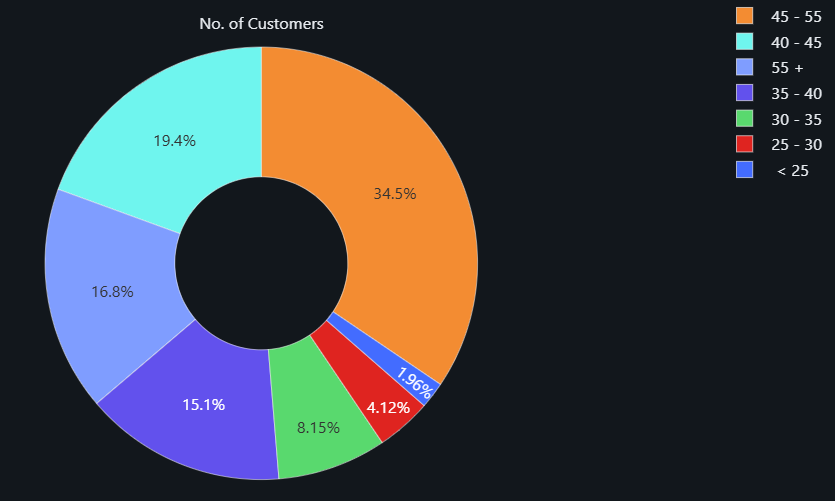
Looking at the distribution across gender for both cohorts – All Customers (Left) and Churned Customers (Right)

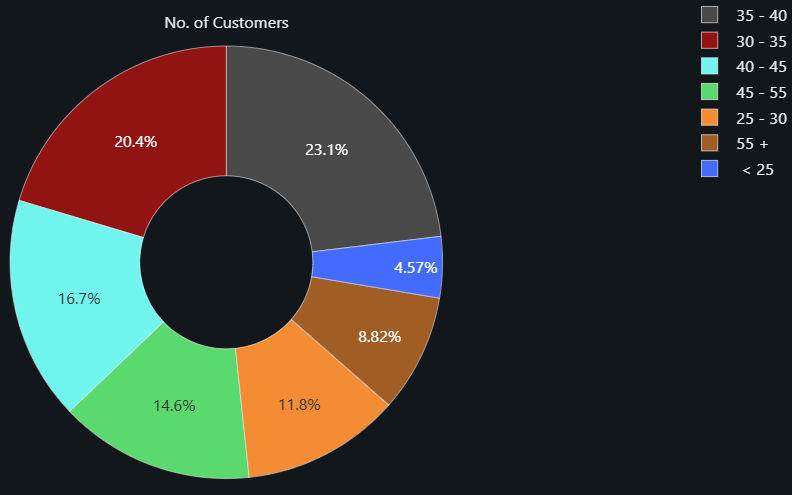


We can see that there is a slightly higher proportion of Female Customers who churned out compared to the Male Customers. The difference is not as noticeable as the previous 2 insights but might be interesting to further investigate upon.

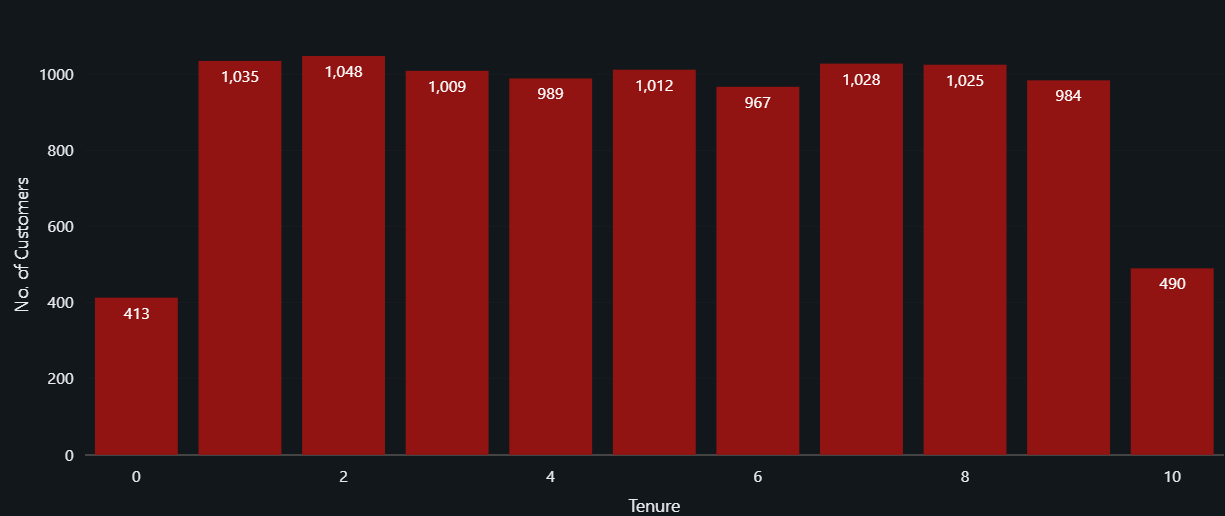
So far, we’ve seen that the customers with the highest likelihood of churn tend to be women from Germany with a Credit Score < 400.

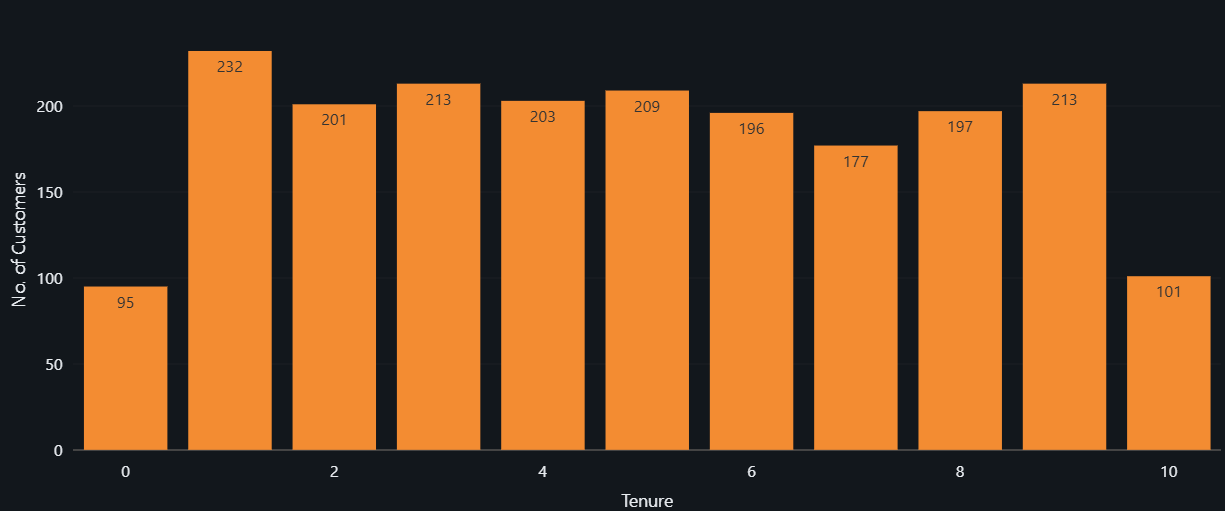
We can further check demographic divides such as age, bank products/services, tenure etc.

Looking at the distribution across age for both cohorts – All Customers (Left) and Churned Customers (Right),

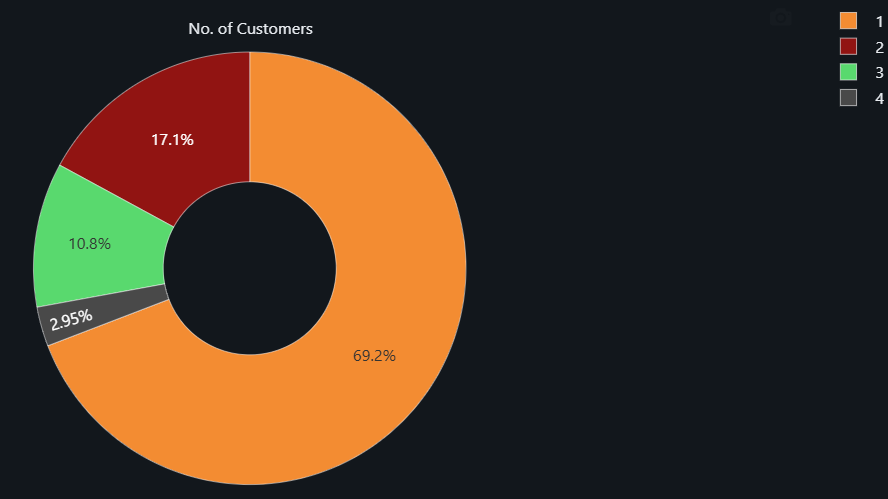


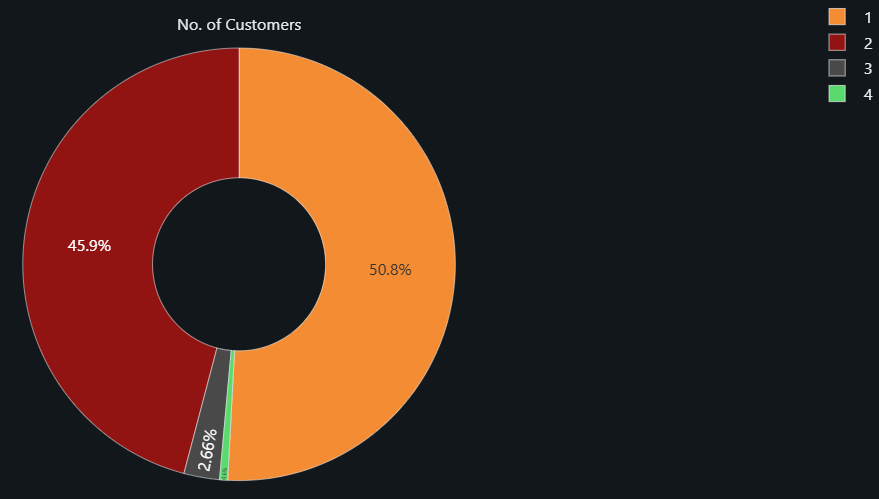
There is a difference in how the customers in both cohorts are distributed across age groups. We can see that the people aged 30-35 are highly unlikely to churn whereas people aged 45 and higher have a larger likelihood of churning.

Looking at the distribution across tenure with the bank for both cohorts – All Customers (Top) and Churned Customers (Bottom), 

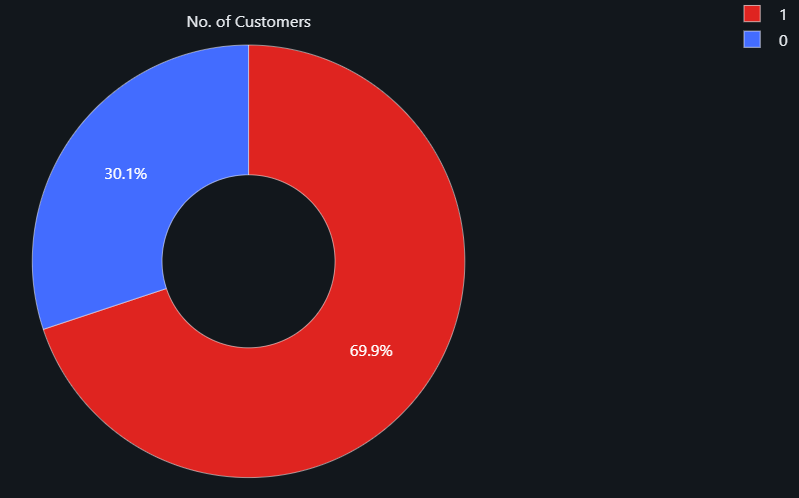


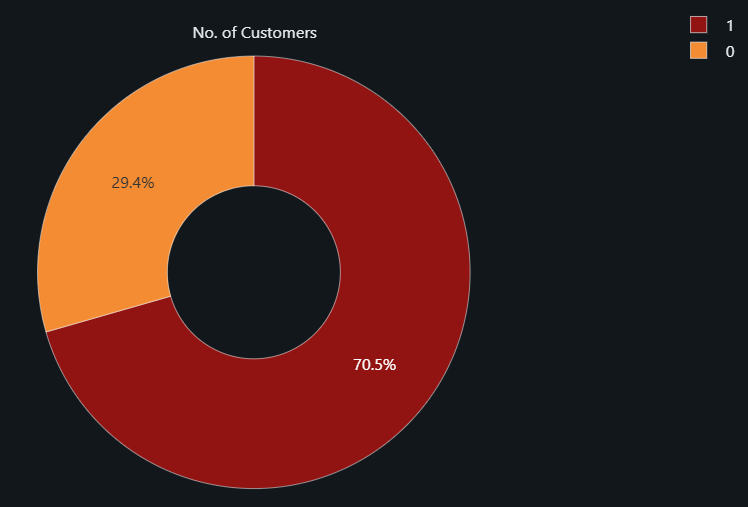
We cannot see any clearly distinguishable patterns based on customer tenure with the bank to identify customers at a higher risk of churn.

Looking at the distribution across number of bank products/services being used for both cohorts – All Customers (Left) and Churned Customers (Right),



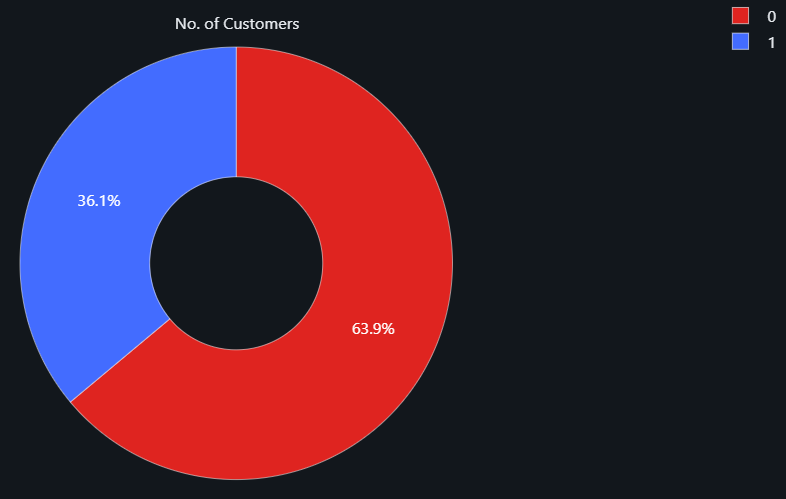
We can clearly see that Customers using more than 3 of the Bank’s products/Services are at a higher likelihood of churn. Despite being <4% of the Bank’s total consumer base, they constitute roughly 14% of the Churned Customers.

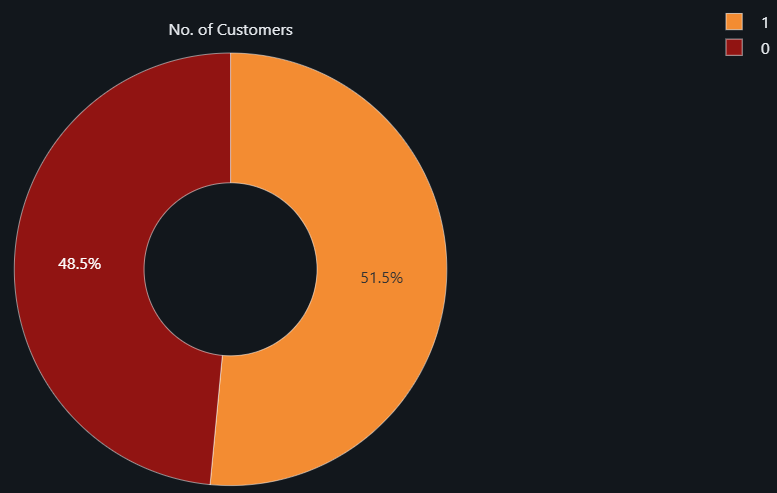
Looking at the distribution based on ownership of a credit card – All Customers (Left) and Churned Customers (Right),



We cannot see any clearly distinguishable patterns based on the customer owning or not owning a credit card.

Lastly, we also have a field indicating whether or not a user is an Active Bank customer. This can help us decide if inactivity can be an early indicator of churn.

Looking at the distribution based on Active Customer Status – All Customers (Left) and Churned Customers (Right),



We can clearly see that Customers who have been inactive in recent times have contributed towards churn in a much higher proportion. This means that Inactivity can indeed be treated as an early indicator of churn.

Based on this EDA, we can conclude the following 2 statements,

* The Bank’s most common customer demographic is a French Man somewhere in the 30-45 age group with a credit card and a credit score of 600 – 700. This customer also uses just 1 of the bank’s products/services.
* The most at-risk demographic, with the highest likelihood of churn is a German Woman somewhere in the 45+ age group with a credit card and a credit score lower than 400. This customer uses more than 2 of the bank’s products/services

Using this information, we can build a rule-based model with if-else conditions that can help us identify such customers so that the bank can investigate their accounts more thoroughly and take pre-emptive rectification actions to minimize churn.

**Estimating Churn – Classification Modelling**

So far we have a rule based prediction in place, however, given the sufficient data volume at hand we can also go about doing some classification modelling to get precise probabilities of how likely a customer is to churn based on their available metadata and demographic information.

The data at hand has a lot of cardinality and too many categorical columns – credit card ownership, active member, country, gender etc. To avoid One Hot Encoding all of these columns and there keeping the number of predictors low and interpretable it is best to use tree based models to classification.

We’ll be using 3 different classification models and compare them on the basis of the ROC-AUC metric.

The 3 models include,

* Decision Tree – A singular tree to deal with out dataset without having to worry about encoding of categorical data. Trees perform well but are hugely outperformed by other more advanced and state of the art models
* Gradient Boosted Trees – GB Trees are a form of a boosting ensemble method. This model makes use of multiple trees and therefore is able to perform better than a singular trees.
* Random Forest – Random Forests are a form of a bagging ensemble method. This model trains multiple trees with a different training method from GB Trees and is also able to outperform singular trees.

Checking the performance of all 3 models on our dataset in predicting churn we observe the following,

|  |  |  |
| --- | --- | --- |
| **Model Name** | **ROC - AUC** | **Output** |
| Decision Tree | 0.5820 |  |
| Gradient Boosted Tree | 0.8689 |  |
| Random Forest | 0.8365 |  |

The theoretical information about the models seems to match up with the performance on the dataset as well. The single Decision Tree does well (0.5820 AUC) but is hugely outperformed by GB Trees and Random Forests, both of which provide a performance of over 0.83.

We can also see that the GB Trees perform better than the Random Forests in this case and although that does seem promising, the performance of both these models after proper hyperparameter fine tuning, cross validation and feature engineering/feature selection can change by leaps and bounds.

All models return a similar hierarchy when it comes to feature importance as well with Age being the most important in all 3 of them and Credit Card ownership being the least important in all three.

An interesting thing to note here is that all the fields that show little to no importance in the models are the ones we couldn’t see any noticeable patterns in during EDA.

In essence we can go ahead and say that where the Rule Based model gives us a solid binary idea about what fields to base our prediction on, these models give us a weighted equation with each fields getting their own importance and corresponding weights.

We now have 3 total classification models 2 out of which are extremely promising and can be built upon further for such use cases with supplemental data, historical information, and other important features.

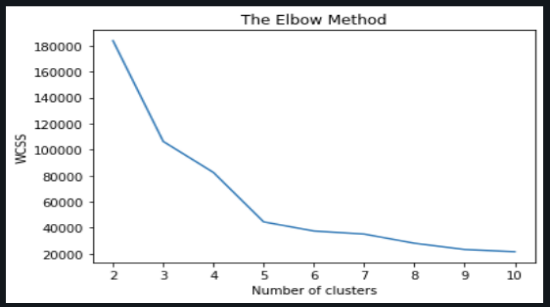
**Identifying High Value Churn – Customer Clustering**

So far, we’ve done 2 things for the bank,

* We’ve identified what their target demographic looks like and which sub-set of this demographic is at the highest likelihood of churn
* We’ve built and provided 3 different classification models to more precisely handle the churn probability and provide meaningful and quantifiable numbers for every customer based on their demographic information and account metadata.

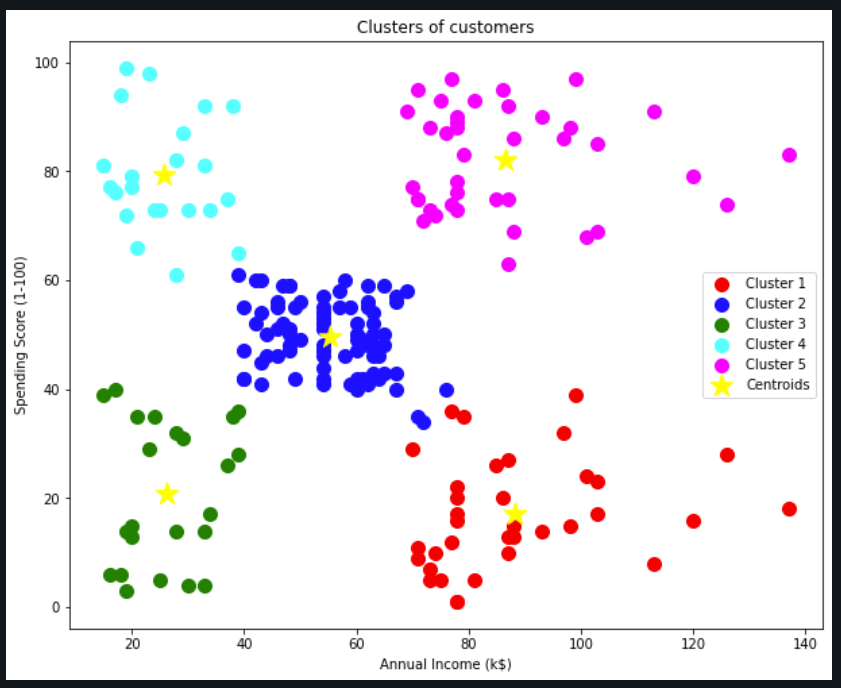
The bank now wants to start addressing the problem of churn and they have 2 different ideas in mind for doing this. On one end they want to run Ad campaigns with existing customers in order to get them to sign up for more services and continue to be active users, therefore reducing their likelihood of churn. At the same time, they also want to reach out to previously churned out customers through a re-entry program to get them to reopen their account with added incentives.

For this part of the project, we’ll look at re-entry program. The bank has identified 200 churned customers for a pilot program of this initiative. However, to do not wish to reach out to all 200 of these customers. They only wish to reach out to customers who it would be profitable to retain at the bank based on their income and spending habits.

In order to do this, we’ll be using a clustering algorithm to group similar customers together based on their income and spending and then pick which clusters would make the most sense to go after through this pilot program.

The first thing to do is, we use the Elbow Method technique to find out the optimal number of clusters for our problem. We do this by changing the number of clusters and measuring the Within Cluster Sum of Squares. The number of clusters beyond which WCSS stops changing significantly is optimal.

Based on the plot of the WCSS v/s Number of Clusters, we can see that we have 5 as the number of optimal clusters. Based on this information, we can go ahead and cluster our datapoints into 5 different clusters. We’ll be using K-Means clustering for this purpose.



Based on these Clusters we can assign the following names to our 5 clusters,

* Cluster 1 – High Income, Low Spending
* Cluster 2 – Medium Income, Medium Spending
* Cluster 3 – Low Income, Low Spending
* Cluster 4 – Low Income, High Spending
* Cluster 5 – High Income, High Spending

From a bank’s point of view, we don’t want anyone who doesn’t have a high income or a high spending at least. This means that customers from Cluster 2 and Cluster 3 should not be considered for our pilot program.

Similarly, Cluster 1 should also be discarded since they do have a high income but not enough of a spending for it to be profitable for the bank.

That leaves us with the 2 target clusters – Cluster 4 and Cluster 5. Cluster 4 is profitable for a bank since they make low income and have a high spending. This means that they would be good candidates for recurring, low value lines of credits or, ideal candidates for credit cards. Cluster 5 is profitable since they qualify for a credit card but are also ideal for high value, long term lines of credit or loans. Therefore, they are the prefect candidates for the re-entry pilot program. Going after these clusters will have the highest return on investment in terms of monetary benefit.

**Customer Retention Ad Campaign – Multi Armed Bandit Problem**

We have now address one part of our problem which is the re-entry program and the ideal churned customers to go after. The second part of minimizing churn is to run a retention program through Ad Campaigns.

The dataset at hand has information about 10 different Ads that were shown to 10,000 customers. Each customer was only shown all 10 ads during an online banking session and their interaction with each Ad is marked with a Boolean variable in data. Out goal is to find out which ads performed the best so that we can cut down on poorly performing ads and invest more on the ones which perform better.

More importantly, since each ad has a cost of running on the web, we want to figure out the most successful ads as quickly as possible to minimize experimentation cost and maximize customer engagement.

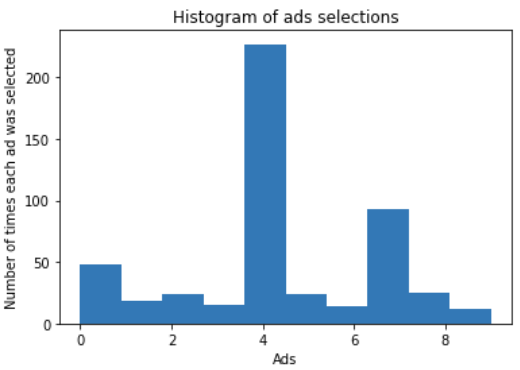
This is what is called a Multi-Armed Bandit problem in Reinforcement Learning and the problem is to figure out the distribution of a dataset while traversing the dataset. Essentially, we want to explore and exploit a dataset at the same time. So in this case, we have 10 different Bandits (Ads 1 through 10) and while we traverse the data for these ads, we simultaneously want to figure out their distribution (Users clicking v/s not clicking) to get the best results for our investment.

We’ll be implementing 2 different algorithms for this purpose to see which one performs the best. The metric of comparison here will be how many iterations it takes for the algorithm to converge to the best performing ad. The 2 algorithms are,

* Upper Confidence Bound – This algorithm starts off by assuming a starting point and an expected value for each Ad and with every iteration, it created an Upper Confidence Bound for the ad. This Confidence Bound decides the probable return from that particular Ad and the Ad with the higher Confidence Bound is chosen.
* Thompson Sampling – This algorithm uses Bayesian Inference to setup a reward system for each Ad at every iteration. At then end of each iteration, the Ad with the highest reward is selected as the best performing ad.

Performance for the Upper Confidence Bound (UCB) Algorithm,

|  |  |
| --- | --- |
| 10,000 Rounds |  |
| 5,000 Rounds |  |
| 1,000 Rounds |  |
| 500 Rounds |  |

We can see that the UCB algorithm does well but needs somewhere between 500 – 1000 rounds before it can determine the best performing ad. That is 500 iterations of all 10 ads running on the web which can be expensive. Although being able to judge the best performing ad in 1000 out of 10,000 round is impressive, the goal is to see if we can do better.

We ran the Thompson Sampling model with 500 rounds and the model was able to converge onto the best performing ad.

The screenshot on the right are the results after running the Thompson Sampling algorithm for 500 rounds.

**Conclusion**

* We were successfully able to determine the Bank’s customer demographic. This information can be used by the bank in their acquisition and outreach programs to get more people to sign up for their services to counter the problem of churn. If rate of acquisition > rate of churn then the bank can continue to make money.
* We were successfully able to identify the most at-risk demographic subset, with the highest likelihood of churn based on EDA which can be used to setup a rule-based model by the bank to pre-emptively flag delinquency and probability of churn.
* We also built and compared 3 different tree-based machine learning models – Decision Trees, Gradient Boosted Trees and Random Forests to predict the likelihood of a customer based on their information. This gives us a comparatively more deterministic approach towards identifying customers at risk of churn.
* We helped identify High Value customer segments for the re-entry initiative’s pilot program. The bank can target the provided segments to re-engage with them and reverse the churn by adding welcome back incentives
* We provided a dynamic solution to the Multi Armed Bandit Problem faced by the Retention Ad Campaign. We can now run experiments with multiple ads and can guage the best performing ad relatively quickly to capitalize on it and avoid wasteful expenses.