Bank Account Retention

Team Members: Michael Claude (01956328), Vivian Kwong (01959505)

Anish Kanade (01957142), Lakshmi Pranathi Vutla (02197113),

Dandu Narasimha Raju(02198757), Sri Charan Tadiparthi(02176302)

University of Massachusetts Boston

Presentation Date: 12/11/24

Overview:

This project aims to construct a model that can predict whether or not a customer will churn (exit) from a bank. We did this by covering three different AI Models (Logistic Regression, Neural Networks, Clustering) and training them on a dataset of customer churn data. The results showed that while the more complex Neural Network model could yield better metrics, these improvements were only marginal as compared to a much simpler Logistic Regression Model. We also noticed key unique features within our dataset through our Clustering models such as Balance and Estimated Salary.

Our final Conclusion was that account holders who are younger, have less money in their accounts, or haven't been with the bank for long have a higher exit rate. The models we tested helped to showcase these results and provide a framework to better predict whether or not a customer will churn and what features to focus on when looking into account retention.

Table of Contents:

1 Introduction

1.1 Background, Motivation, and Problem Statement

The problem statement for this project was that we wanted to predict whether or not a customer would stay with the bank or leave based on customer retention data. This is vital information for a Banking business and could give insights into customer trends. It could also be used to track Bank performance as an increase in churning rates could be indicative of poor management.

Customer retention is also helpful financially as it is more cost effective to retain an already existing customer than to acquire and onboard a new one. Figuring out factors can lead to a customer exiting would be beneficial in better accommodating and understanding these groups.

1.2 Objectives

Below we will list a few of our primary objectives with this project:

- Develop an AI Models to predict customer exit rates
- Discover what features play large roles within customer churning
- Refactor data to be readable and understandable

2 Literature Review

2.1 Related Works

On the Kaggle page for our dataset, there are around 75 other implementations of churn predictions. [1]

3 Methodology

3.1 Data Description and Source

The data we used was from a Kaggle data set[2] which contained different bank accounts and their churn status. This data can be used to detect trends within customer churning as well as make future predictions. Below are the given features for a customer entry:

- **CustomerId**: Contains the customer's identification number.
- **Surname**: The surname of a customer.
- **Credit Score**: Current credit score of the customer.
- **Geography**: A customer’s geographic location.
- **Gender**: Gender of customer.
- **Age**: Age of customer
- **Tenure**: Number of years the customer has been with the bank.
- **Balance**: Current balance within the customer’s bank account.
- **Estimated Salary**: Estimated salary of the customer.

Inside of our dataset, around ~8000 customers have not churned while ~2000 have churned.

3.1.1 Data Preprocessing

For Data Preprocessing, we detected null columns within our dataset. Most importantly, we dropped the ‘Exited’ and ‘CustomerId’ features. ‘Exited’ was dropped from our training dataset as we would use this value as our target variables. ‘CustomerId’ was dropped due to being non-predictive of churn status.

3.2 Model Selection

We chose three different models to do testing with: Logistic Regression, Neural Networks, and Clustering.

3.2.1 Justification

Below we will discuss the justifications for each model selected:

- Logistic Regression: Logistic Regression was chosen as a simple baseline model as it is relatively simple and easy to work with. This model also works very well with binary Classification problems such as ours.
- Neural Networks: Neural Networks was a more complex model we used to see if higher complexity would benefit our case to justify it. Neural Networks also work well within Classification problems making it a prime candidate.
- Clustering: Clustering was a third option we chose so to take a slightly different route from the other two. Clustering models work with different groups and can be used to find unique associates between said groups. In our case where we have multiple different features in which to group customers, Clustering could be used to predict churn rates as well.

3.3 Implementation Details

3.3.1 Logistic Regression

Logistic Regression being our simple baseline means the implementation was relatively simple as well. We first split our dataset into our training and testing sets. Then, we standardized our data to better accommodate the model. Next, the model was created and fitted to our data. Finally, we ran predictions on the test set.

To analyze our model, we generated its Confusion Matrix, ROC Curve, Logistic Loss, and learning curve.

3.3.2 Neural Networks

The setup for our Neural Network model implementation is similar to that of our Logistic Regression model. We start by splitting our data into train and test sets and then standardizing them for the model. Our Neural Network model is then made with one input layer with 64 nodes, one hidden layer with 32 nodes, and then one output layer with 1 node as our binary classification. The input and hidden layers both use ReLu as the activation function with the output layer using Sigmoid.

Our model is compiled with an Adam Optimizer, loss measured with binary cross-entropy, and accuracy metrics. We then train our model for a certain number of epochs and finally evaluate it on the test set.

For analysis, we generated the Confusion Matrix, plotted our training and validation loss over epochs, and plotted our training and validation accuracy over epochs.

3.3.3 Clustering

We applied clustering to individual features (e.g., geography, gender, credit score, age, tenure, balance, and estimated salary) to identify patterns that may influence customer churn. By isolating each feature, our goal was to understand its specific contribution to predicting churn.

For categorical features like geography and gender, we first encoded them into numerical values and stored the mappings for interpretability. For numerical features, we standardized the data to ensure uniform scaling, which is crucial for distance-based methods like KMeans clustering.

Alongside KMeans, we explored simpler clustering techniques, such as binning or grouping based on feature thresholds, to segment data into meaningful clusters. This approach allowed us to gain initial insights into the feature distributions without relying on complex distance-based methods.

We visualized these clusters using descriptive statistics and boxplots to compare their properties, focusing on key differences related to churn. For instance, we examined how average balance levels or tenure varied across clusters and how these patterns correlated with churn rates. This analysis provided actionable insights into the characteristics of customer groups at higher risk of leaving the bank.

3.4 Tools and Libraries

For Fine Tuning, we used:

- Pytorch
- Tensorflow

For Data Exploration and Visualization, we used:

- Matplotlib
- Seaborn

For Model Testing and Analysis, we used:

- Scikit-learn
- Pandas

For Collaboration, we used:

- Google Colab
- Google Slides/Docs
- Whatsapp

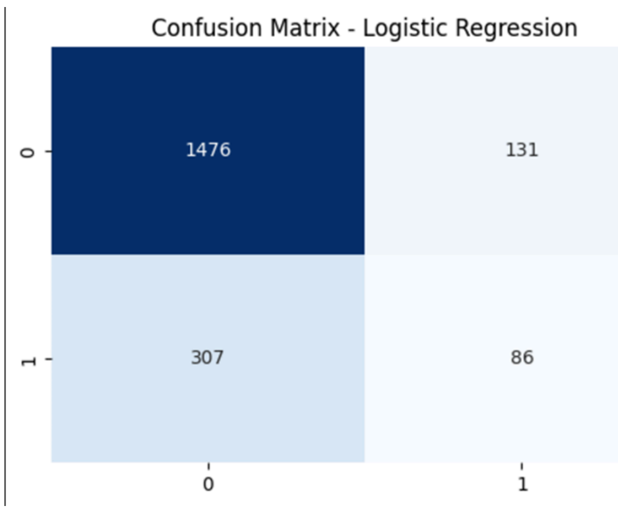
4 Experiments and Results

4.1 Experiment Setup

For our testing, we trained each model on the same data to test their accuracy against each other. This involved the previously mentioned implementation and analysis detailed inside of section 3.3. Inside of this section, we will go over more of the specific details, numbers, and results.

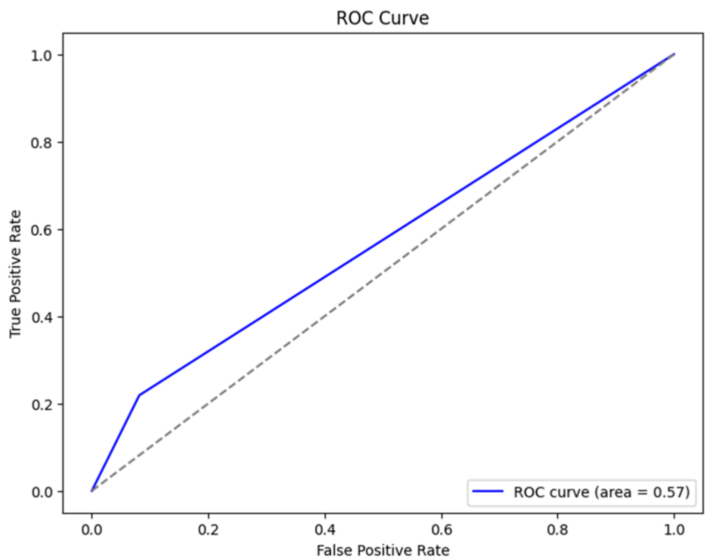
4.2 Performance Metrics and Analysis

4.2.1 Logistic Regression Performance Metrics

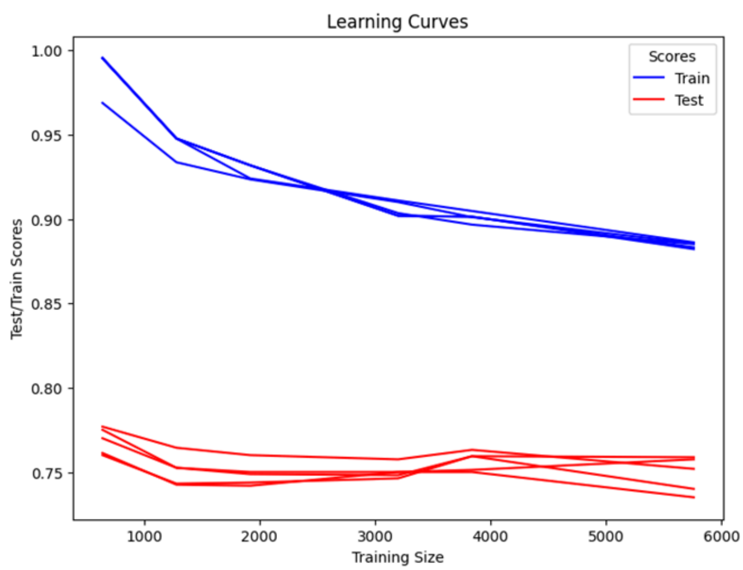


This is the Confusion Matrix for Logistic Regression. 0 here means the customer has not exited and 1 means the customer has exited. Most of the customers would fall within our True Negative case which aligns with our data since a majority of our customers have not exited and the Logistic Regression model correctly predicts this.

One area for concern is our False Positive rate which could be an area for future improvement and focus.

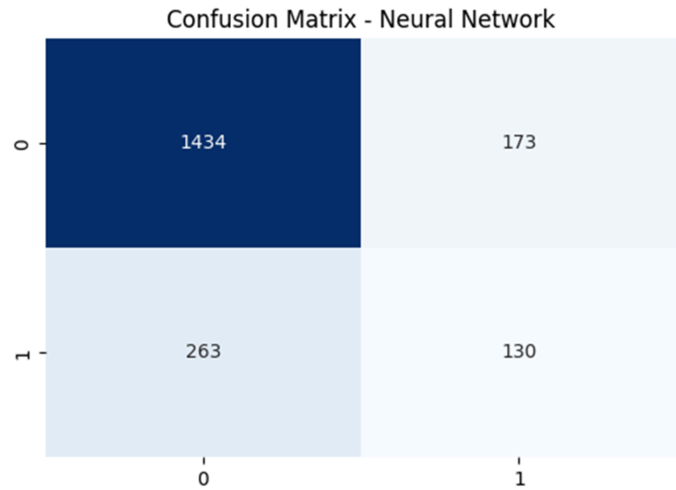


Next is our ROC Curve for Logistic Regression which shows the function of our False Positive rate (People flagged as churned who did not churn) and our True Positive rate (People flagged as churned who did churn).

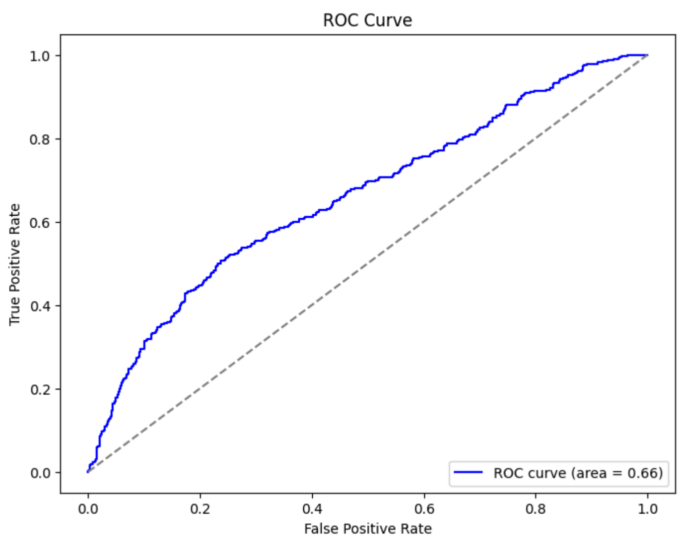


Lastly, for Linear Regression we looked into the Learning Curves which cover our Train and Test scores over an increasing dataset size. Our train set is a subset of the dataset known to the model while training and the test set is a subset unknown to the model while training. The accuracy of the train set starts to dip the larger our dataset size gets however our test accuracy rate remains relatively constant. This means our model may require additional optimizing and testing to get our Test score to begin to go up as we increase our dataset size.

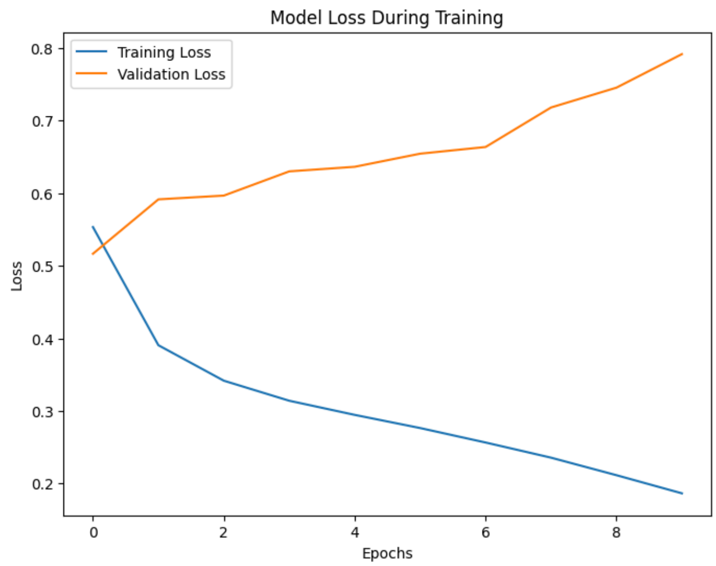
4.2.2 Neural Networks Performance Metrics



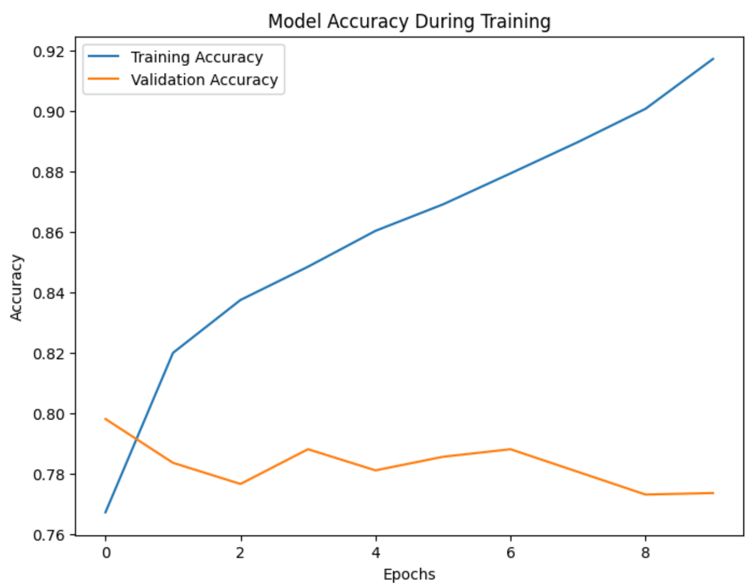
This is the Confusion Matrix for Neural Networks. The labels are the same as Logistic Regression with 0 being Not Exited and 1 being Exited. The True Positives and Negatives have good accuracy here though also similar to Logistic Regression having a noticeable issue when it comes to False Positives.



Next is our ROC Curve for Neural Networks which shows the function of our False Positive rate (People flagged as churned who did not churn) and our True Positive rate (People flagged as churned who did churn).



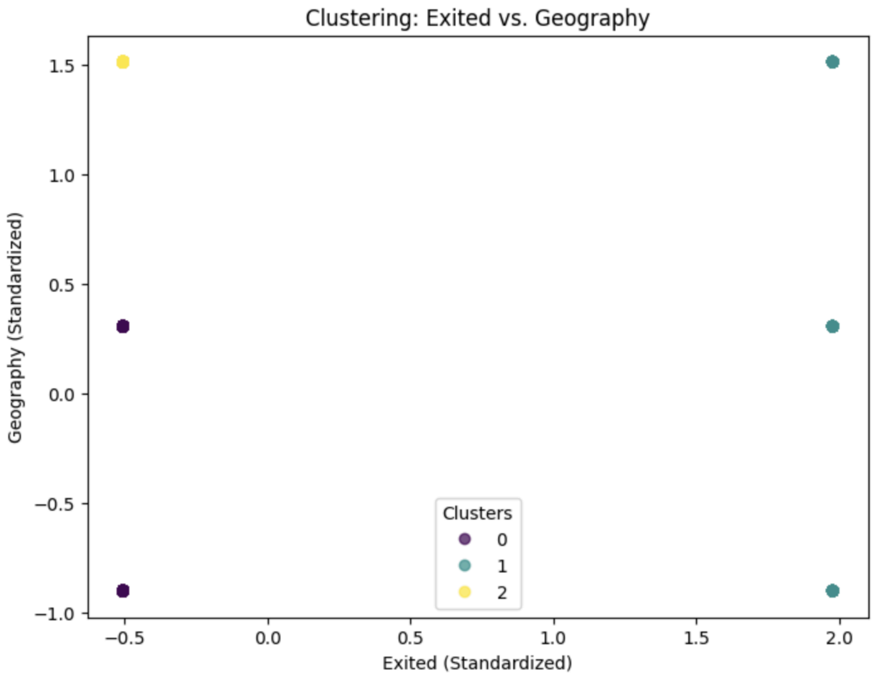
This graph shows our Training and Validation loss as we test with different epochs. This pattern shows that our model is possibly overfitting[4]. We could optimize this more by lowering the size of our Neural Network model.



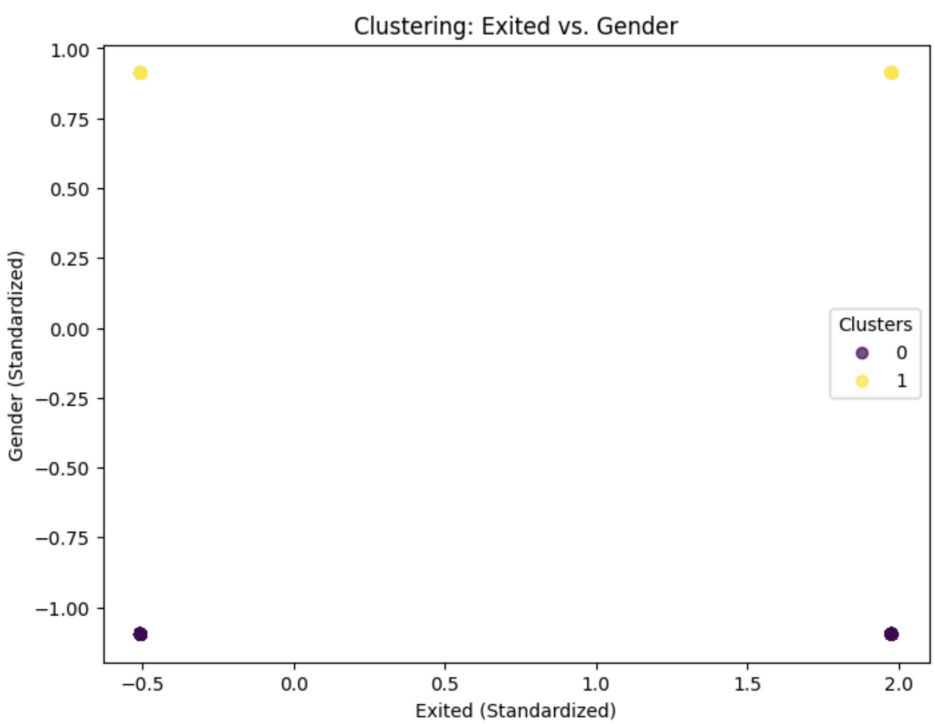
This graph displays our Training and Validation accuracy over epochs. Once again, this could be a sign of overfitting as our model becomes very familiar with our training data however loses accuracy with validation data.

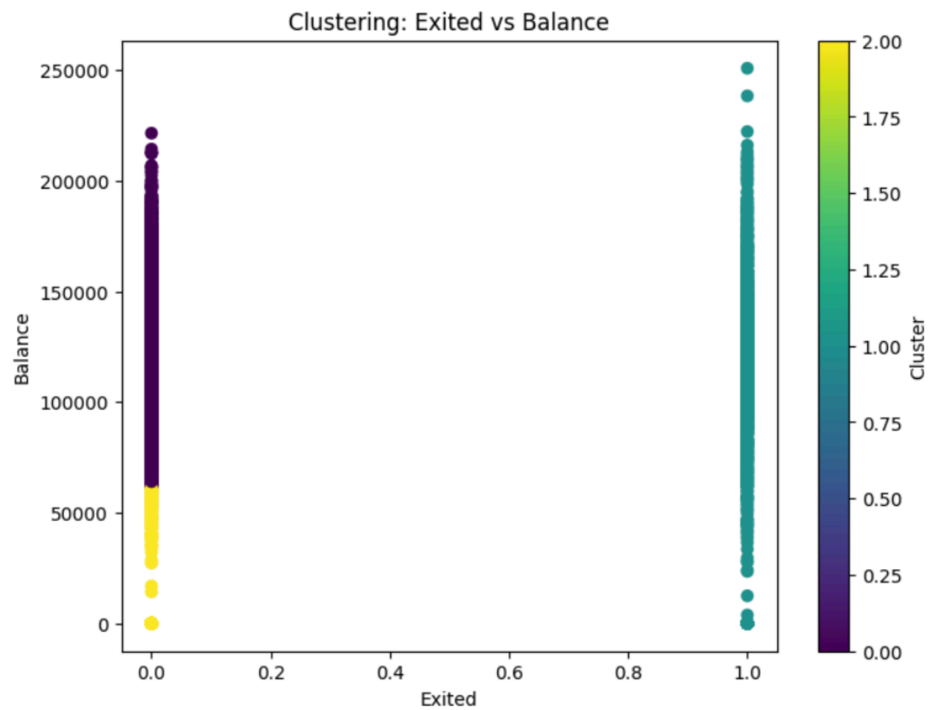
4.2.3 Clustering Performance Metrics

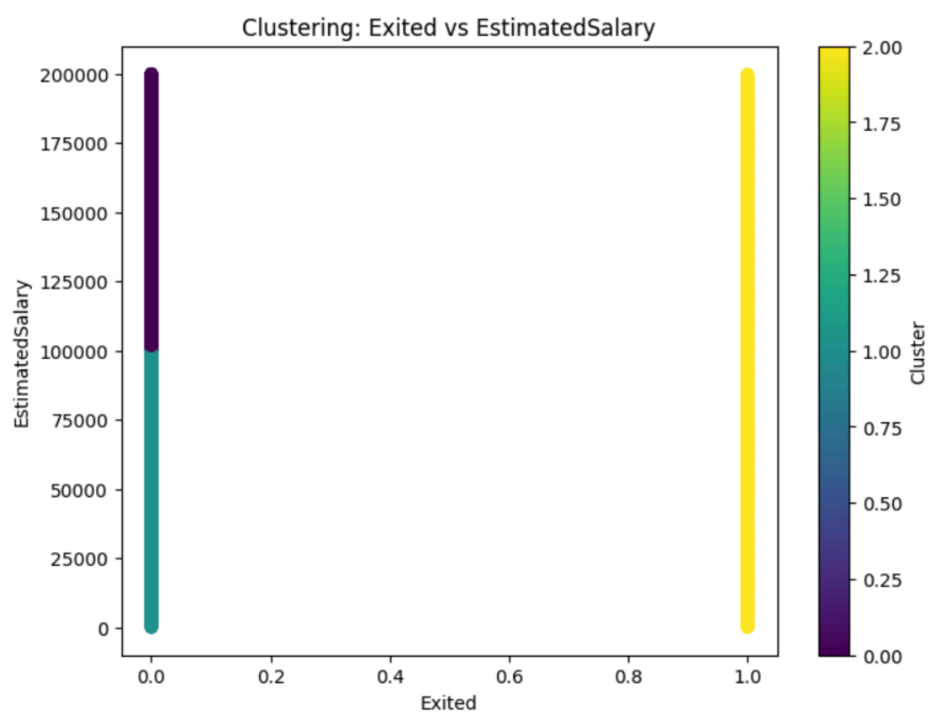
For clustering based on geography, a plot of the clusters shows that there are distinct clusters based on whether the customer is from France (encoded as 0), Germany (encoded as 1), or Spain (encoded as 2).

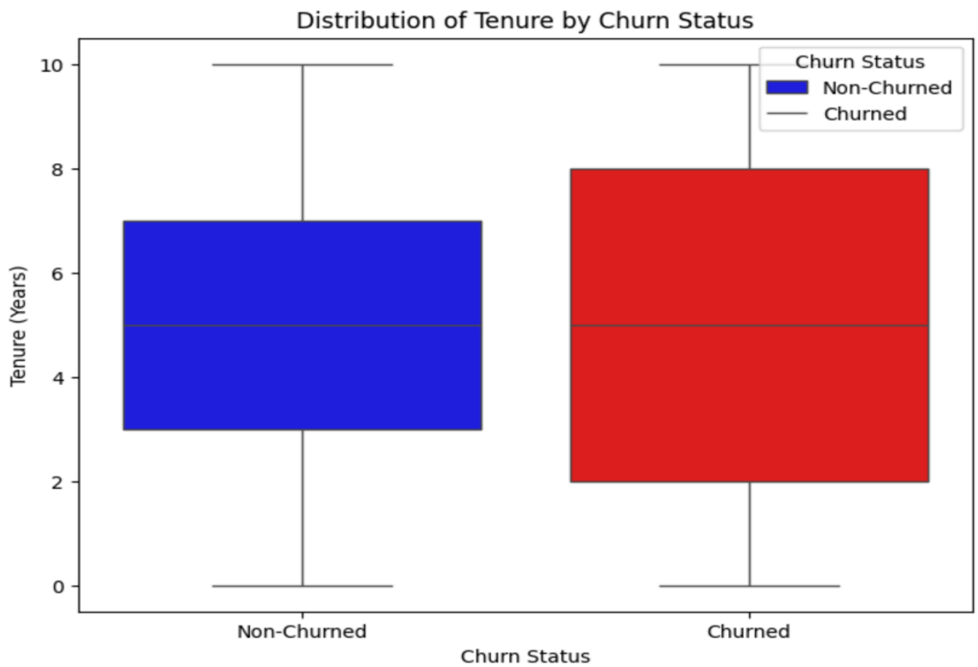


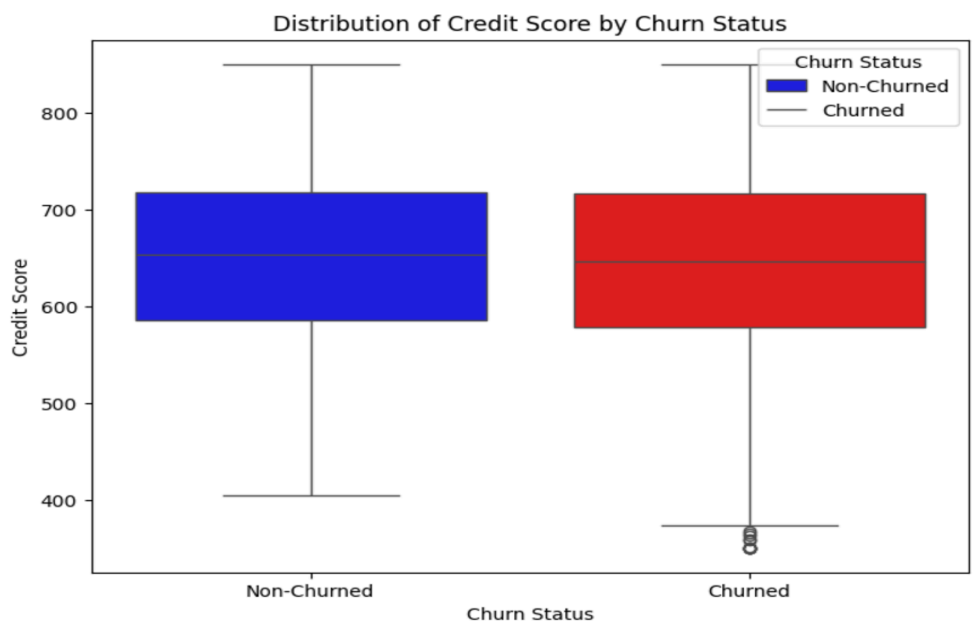
For clustering based on gender, a plot of the clusters shows that there are distinct clusters based on whether the customer is a female (encoded as 0), or a male (encoded as 1).




Clustering: Exited vs Balance
Balance
Exited
Cluster
250000
200000
150000
100000
50000
0
0.0
0.2
0.4
0.6
0.8
1.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00


Clustering: Exited vs EstimatedSalary
EstimatedSalary
Exited
Cluster
200000
175000
150000
125000
100000
75000
50000
25000
0
0.0
0.2
0.4
0.6
0.8
1.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00





4.3 Results

4.3.1 Logistic Regression Results

The following are the results of our Logistic Regression Testing:

- Precision: 83% (Not Exited), 40% (Exited)
- Recall: 92% (Not Exited), 22% (Exited)
- F1-Score: 87% (Not Exited), 28% (Exited)
- Accuracy: 78%

4.3.2 Neural Networks Results

The following are the results of our Neural Network Testing:

- Precision: 85% (Not Exited), 43% (Exited)
- Recall: 89% (Not Exited), 33% (Exited)
- F1-Score: 87% (Not Exited), 37% (Exited)
- Accuracy: 78.2%

4.3.3 Clustering Results

Results for Cluster analysis for Exited vs. Gender:

Note: 0 is France, 1 is Germany, 2 is Spain

```
                       CreditScore  Geography  Gender        Age    Tenure  \
Cluster_Exited_Gender
0                       650.831389   0.742021     0.0  39.238389  4.966102
1                       650.276892   0.749863     1.0  38.658237  5.051677

                            Balance  NumOfProducts  HasCrCard  IsActiveMember
Cluster_Exited_Gender
0                      75659.369139       1.544134   0.702619        0.502751
1                      77173.974506       1.518600   0.707898        0.525380

                       EstimatedSalary    Exited  Cluster_Exited_Geography
Cluster_Exited_Gender
0                        100601.541382  0.250715                  0.628439
1                         99664.576931  0.164559                  0.606560
```

Results for Cluster analysis for Exited vs. Geography:

Note: 0 is female, 1 is male

```
                         CreditScore  Geography    Gender        Age  \
Cluster_Exited_Geography
0                         651.765384   0.287337  0.568401  37.257671
1                         645.351497   0.805106  0.440844  44.837997
2                         652.104167   2.000000  0.584302  37.839147

                           Tenure        Balance  NumOfProducts  HasCrCard
Cluster_Exited_Geography
0                         5.007289  77317.386133       1.544160   0.709951
1                         4.932744  91108.539337       1.475209   0.699067
2                         5.107558  59678.070470       1.544574   0.699128

                         IsActiveMember  EstimatedSalary  Exited  \
Cluster_Exited_Geography
0                              0.550093    100135.874263     0.0
1                              0.360825    101465.677531     1.0
2                              0.567345     98602.369864     0.0

                         Cluster_Exited_Gender
Cluster_Exited_Geography
0                                     0.568401
1                                     0.440844
2                                     0.584302
```

Results for Cluster analysis for Exited vs. Balance:

```
Analysis for 'Exited' vs 'Balance'

Cluster-wise Average Values for 'Exited' vs 'Balance':
                        Exited        Balance
Cluster_Exited_Balance
0                          0.0  121414.606229
1                          1.0   91108.539337
2                          0.0    2020.363731

Number of Data Points in Each Cluster (Exited vs Balance):
Cluster_Exited_Balance
0    4717
2    3246
1    2037
Name: count, dtype: int64
```

Results for Cluster analysis for Exited vs. Estimated Salary:

```
Analysis for 'Exited' vs 'EstimatedSalary'

Cluster-wise Average Values for 'Exited' vs 'EstimatedSalary':
                       Exited  EstimatedSalary
Cluster_Exited_Salary
0                         0.0    149474.052023
1                         0.0     50413.255980
2                         1.0    101465.677531

Number of Data Points in Each Cluster (Exited vs EstimatedSalary):
Cluster_Exited_Salary
1    3998
0    3965
2    2037
Name: count, dtype: int64
```

Results for Cluster analysis for Exited vs. Tenure:

```
Performing clustering for 'Exited' vs 'Tenure'...
Analysis for 'Exited' vs 'Tenure'

Cluster-wise Average Values for 'Exited' vs 'Tenure':
                         Exited    Tenure
Cluster_Exited_Tenure
0                      0.193548  6.005321
1                      0.210058  2.250556
2                      0.204482  8.785914

Number of Data Points in Each Cluster (Exited vs Tenure):
Cluster_Exited_Tenure
1    4494
0    3007
2    2499
Name: count, dtype: int64
```

Results for Cluster analysis for Exited vs. CreditScore:

```
Performing clustering for 'Exited' vs 'CreditScore'...
Analysis for 'Exited' vs 'CreditScore'

Cluster-wise Average Values for 'Exited' vs 'CreditScore':
                              Exited  CreditScore
Cluster_Exited_CreditScore
0                           0.218878   529.625418
1                           0.197613   766.316251
2                           0.198430   649.513677

Number of Data Points in Each Cluster (Exited vs CreditScore):
Cluster_Exited_CreditScore
2    4460
1    2849
0    2691
Name: count, dtype: int64
```

4.5 Analysis

4.5.1 Logistic Regression and Neural Network Comparison Analysis

The Logistic Regression model and the Neural Network models were the two models we wanted to compare. As mentioned before, Logistic Regression models are simple and effective for binary classification tasks. Neural Networks are much more complex and require more tuning, however are great at discovering non-linear patterns.

While Neural Networks eased out ahead of Logistic Regression, it was only by a few percent points (<3%). With our current dataset, Logistic Regression seems better fit as we can try to avoid overfitting and have an easier time tuning due to its simplicity.

4.5.2 Clustering Analysis

Clustering based on geography reveals that customers from Germany are more likely to churn in comparison to customers from France and Spain. Further analysis of customers from these different geographies show that customers from Germany are less active at their bank and have higher balances than customers from France and Spain.

Clustering by gender shows that female customers are less likely to churn compared to male customers. Additionally, female customers exhibit higher levels of activity at the bank, despite other features, such as account balances and tenure, being similar across the clusters.

Clustering by balance and EstimatedSalary shows that customers with high balances and high salaries are more likely to remain loyal (Cluster 0 in 'Balance'). customers with moderate balance and moderate salary are at risk of leaving (Cluster 1 in 'Balance', Cluster 2 in Estimated Salary).

Clustering by Tenure shows that customers with shorter tenure (Cluster 1) are more likely to leave than those with longer tenure (Clusters 0 and 2). This emphasizes the value of early-stage retention strategies. customers with longer tenure (Cluster 2) have a significant churn rate (~20.45%), indicating that even long-term customers need engagement to stay loyal.

Clustering by CreditScores shows that higher credit scores may be associated with slightly lower exit rates, suggesting that creditworthiness is a weak predictor of churn. However, since the variation in Expected rates is small, other factors beyond CreditScore might play a more significant role in determining customer churn.

5 Discussion

5.1 Insights

Understanding the features and variables that influence a customer's decision to leave a bank helps identify areas where the bank can allocate resources for effective retention strategies.

From our clustering analysis, we observed that German customers and male customers are more likely to churn. To address the geographical trend, the bank might focus on understanding and appealing to the specific values and preferences of German customers. For instance, the lower activity levels among German customers, compared to those in France and Spain, suggest an opportunity to engage this demographic more actively.

Similarly, the bank could analyze why female customers are more likely to stay and apply these insights to develop strategies that encourage male customers to remain loyal. The observed correlation between lower activity levels and churn, especially among German and male customers, underscores the importance of fostering higher customer engagement to enhance retention.

Key Insights for Retention Strategies:

1. **Personalized Banking:** The clustering analysis revealed that customers from Germany and males are at higher risk of churn. Developing personalized campaigns targeting these demographics could improve retention.
2. **Engagement Programs:** Offering engagement incentives, such as discounts on fees, cashback offers, or tailored financial planning services, could help mitigate churn.
3. **Retention of New Customers:** Since shorter tenure is linked to churn, focus on onboarding and support programs within the first two years of customer acquisition.

5.2 Limitations and Roadblocks

Listed below are the limitations and roadblocks we ran into while working on this project:

- Learning: We went into this project with very little knowledge of the AI model development field. An initial hurdle was getting comfortable with the libraries, technologies, and terminology needed to get things working.
- Model Selection: As mentioned with Learning, it took us a while to search and find models that would work best for our case.
- Testing/Validation Accuracy: As discussed inside of the analysis, our models have issues when it comes to our validation. With larger dataset sizes and over more epochs, our models grow less and less accurate. For the Neural Network implementation, this can be a sign of overfitting.
- Automation: Currently, insights from Clustering has to be manually applied to the Classification AI Models. It would be ideal if there was some automated process.

5.3 Future Work

In the future, it would be helpful to train our models on more data, such as customers from more countries around the world.

5.3.1 Current Issues to Fix

As mentioned inside of the analysis, we have a few problems with our models. Our Logistic Regression requires a bit more tuning as to increase our Testing Accuracy after training with more and more data. Our Neural Network needs to sort out our overfitting problem by shrinking our Neural Network model size and spend more time tuning it.

Additionally, if we were to acquire more data, we could further our testing with the Neural Network model. This could allow us to see a possible wider gap between Logistic Regression and the Neural Network models as the dataset would grow larger and more complex.

Lastly, we would want to have some automated communication between our Clustering and Binary Classification models. This would allow for stronger predictions backed by our Clustering model.

6 Conclusion

6.1 Summary

In this project, we successfully developed and compared three AI models—Logistic Regression, Neural Networks, and Clustering—to predict customer churn for a bank. Through extensive analysis, we identified critical features influencing churn, such as customer age, tenure, balance, and geographic location.

The Logistic Regression model served as a reliable baseline, demonstrating simplicity and ease of use, while the Neural Network offered slightly better performance, though at the cost of complexity and risk of overfitting. Clustering models provided additional insights into customer demographics and behavior, revealing patterns like higher churn rates among German and male customers.

Key findings include:

1. Younger customers, those with lower balances, and shorter tenure are more likely to churn.
2. Customers from Germany and male customers exhibit higher churn tendencies.
3. Customer engagement is a vital factor in retention, with more active customers showing significantly lower churn rates.

6.2 Implications

The results emphasize the importance of targeted retention strategies, such as personalized outreach, loyalty programs, and tailored banking services. Understanding the factors contributing to churn enables banks to improve their resource allocation and retain valuable customers.

6.3 Future Directions

The project highlights several opportunities for future work:

1. **Improved Model Performance:** Refining the Logistic Regression and Neural Network models to better handle overfitting and validation accuracy issues.
2. **Broader Datasets:** Incorporating more diverse datasets from global banking sectors for generalized predictions.
3. **Automated Insights:** Integrating clustering insights directly into predictive models to enhance accuracy.
4. **Long-Term Monitoring:** Continuously monitoring customer behavior post-prediction to adapt strategies dynamically.

By combining these advancements, the bank can further strengthen its ability to predict and reduce customer churn, ultimately fostering a more loyal and satisfied customer base.

7 References

1. https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers/code
2. https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers/data
3. https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html
4. https://stackoverflow.com/questions/71260933/is-it-normal-for-validation-loss-to-increase-instead-of-decrease
5. https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html
6. https://scikit-learn.org/1.5/modules/neural_networks_supervised.html
7. https://pandas.pydata.org/
8. https://scikit-learn.org/stable/
9. https://seaborn.pydata.org/