# Project Report: Customer Churn Prediction for Banks

**Prepared by:** Sourav Ghosh

## Table of Contents

## 1. Project Goal and Data

Hey there! So, the main goal of this "Churntastic Customer Retention" project is pretty simple: we want to build a cool machine learning model that can guess which bank customers are thinking about leaving us. If we can spot those folks early, we can jump in with some helpful stuff to keep them around. That means happier customers, more value for the bank, and steady growth – win-win!

We got our data from Kaggle, which is a neat place for datasets. It's called "DATA_SET(Churn_prediction).csv" and it's packed with all sorts of info about our customers – like where they live, how old they are, their account details, and what they usually do with their money. Plus, it tells us if they actually left the bank or not (that's our "Exited" column!).

## 2. Data Preprocessing Summary

Before we could get our models learning, we had to do some house cleaning on our data. Here's a quick rundown of what we did:

- **First Look**: We checked out the dataset's size (10,002 rows, 14 columns) and just got a general feel for it.
- **Missing Stuff**: Some rows had gaps in Geography, Age, HasCrCard, and IsActiveMember, so we just took those out. That left us with 9,996 good entries!
- **Double Trouble**: Found some duplicate rows and kicked them out too, making sure our data was super clean.
- **No Need for These**: Columns like RowNumber, CustomerId, and Surname weren't really going to help us predict churn, so we said goodbye to them.
- **Slimming Down**: We tweaked the data types for columns like Age, Tenure,

NumOfProducts, HasCrCard, IsActiveMember, and Exited to int32. This just makes the dataset a bit lighter and faster to work with.

- **New Features!**: We got a bit creative and made new categories for CreditScore and Age called CreditScoreGroup and AgeGroup. Then, we didn't need the old CreditScore and Age columns anymore.
- **Categorical Makeover**: For things like HasCrCard, IsActiveMember, Geography, and Gender, we used something called "one-hot encoding" to turn them into numbers our models could understand. And for our new CreditScoreGroup and AgeGroup, we used "ordinal encoding."
- **Scaling Up (and Down!)**: To make sure all our numbers played fair, we scaled Balance and EstimatedSalary using StandardScaler, and Tenure and NumOfProducts using MinMaxScaler.
- **Balancing Act**: Our "Exited" column was a bit lopsided (way more people stayed than left!). So, we used a technique called SMOTE to balance things out, making sure our model got a fair shake at learning about churners.
- **Picking the Best**: After some tests (Chi-squared and ANOVA F-tests), we realized Tenure and EstimatedSalary weren't super important on their own. So, we tried a version of our data without them (X_resampled_selected) to see if it made a difference.

## 3. Exploratory Data Analysis (EDA) Summary

Diving into the data was a lot of fun! We found some cool stuff about how everything was spread out, how things related to each other, and some hints about why customers might leave:

- **Just the Numbers**: We looked at the average, min, and max for numbers like Age, Balance, and EstimatedSalary. For categories like Geography and Gender, we saw how many customers fell into each group – turns out, most of our customers are in France!
- **One by One**: We made some charts (histograms) to see how numbers like Age and Balance were distributed. Age was a bit skewed, and Balance had two main peaks. For categories, we used bar charts to confirm our geographical findings and see the breakdown of genders, credit card holders, and active members.
- **Two by Two**: When we compared numbers against the Exited column, it looked like older customers and those with bigger bank balances might be more likely to churn. For categories, our bar charts showed that customers in Germany, those with 3 or 4 products, and inactive members were more prone to leaving. Yikes!
- **Churn Check**: A simple bar chart of our Exited variable clearly showed a big imbalance – way more customers stayed than left.

- **Correlation Corner**: We used a heatmap to see how our numerical features were related. It showed Age had a positive link with Exited, and Balance had a weaker positive link.
- **Spotting Oddballs**: Some box plots pointed out a few unusual values (outliers) in CreditScore and Age.

## 4. Model Training and Evaluation Summary

Time for the main event! We trained a few different machine learning models to predict churn:

- **Logistic Regression**: This was our starting point, a pretty straightforward model. We fine-tuned it using GridSearchCV on both our full and "selected features" datasets.
- **Random Forest**: This one's like a team of decision-makers. We trained simple versions first, then really dialed them in with GridSearchCV on both datasets.
- **XGBoost**: This is a super powerful boosting model! Again, we started simple and then fine-tuned it with GridSearchCV on both datasets.
- **Artificial Neural Network (ANN)**: Our deep learning contender! We used some fancy tricks like weight initialization, batch normalization, and L2 regularization to train these on both datasets.

We judged how well these models did based on their accuracy. Here's a little table showing how they stacked up:

| Model | All Features Accuracy | Selected Features Accuracy |
|---|---|---|
| Logistic Regression | 0.6757 | 0.6772 |
| Random Forest | 0.8879 | 0.8572 |
| XGBoost | 0.8920 | 0.8577 |
| ANN (with L2 Regularization) | 0.8134 | 0.7905 |

## 5. Interpretation of Best Model and Key Findings

So, who's the winner? Drumroll, please... the **XGBoost model trained on all features** took the crown with the highest accuracy (89.20%)! This tells us that this "boosting" approach, using all the features we preprocessed and engineered, was the best at figuring out who's likely to churn. Awesome!

Now, understanding exactly *why* XGBoost makes its predictions can be a bit tricky since it's a complex model. But based on our earlier data exploration and feature tests, we can guess that these factors are probably the biggest drivers of churn:

- **Geography**: Customers in Germany seemed to be more likely to pack up and leave.
- **Age**: Older customers appeared to be more prone to churning.
- **Balance**: Folks with bigger bank balances also seemed to be more likely to exit.
- **NumOfProducts**: This is a big one! Customers with 3 or 4 products had a significantly higher churn rate. That's something to look into!
- **IsActiveMember**: If a customer wasn't active, they were more likely to churn than those who were. Makes sense, right?
- **Gender**: There was a noticeable difference in churn rates between genders.
- **CreditScoreGroup**: Customers with lower credit scores (our "Poor" group) were more likely to churn compared to those with higher scores.

Even though Tenure and EstimatedSalary didn't seem super important on their own, including them in the full dataset actually helped our tree-based models (Random Forest and XGBoost) and ANNs perform better. This means they're probably working together with other features in interesting ways!

## 6. What Else Could We Do?

If we wanted to really push this project further, here are some cool things we could try:

- **More Models!**: We could check out other strong models like LightGBM or CatBoost. They're often fantastic for this kind of data.
- **Feature Fun**: Let's get even more creative with our features! Maybe we can build some super complex ones or see how existing ones interact to really boost performance.
- **Imbalance Remix**: We used SMOTE, but there are other ways to handle that lopsided data. We could try different oversampling or undersampling tricks, or focus on metrics like F1-score or AUC-ROC, which are better for imbalanced situations.
- **Hyperparameter Deep Dive**: We did some tuning, but we could go even deeper with techniques like RandomizedSearchCV or even Bayesian Optimization to find the absolute best settings for our top models.
- **Team Up Models**: How about combining the predictions from our best models? That often leads to even better and more reliable results!
- **Explain Yourself, Model!**: For complex models like XGBoost, it's tough to see inside. We could use tools like SHAP values or LIME to really understand *why* the

model is predicting that a specific customer might churn. That's super helpful for taking action!

- **Automate Everything!**: To make this project super easy to repeat and use in the real world, we could build a full "predictive pipeline" that handles everything from cleaning the data to spitting out predictions automatically.

## 7. How Can We Keep Our Customers?

Alright, based on all our findings, here are some simple, actionable ideas for keeping our customers happy and with us:

- **Spot the "Runners"**: Our analysis clearly showed who's most likely to leave – customers in Germany, older folks, those with big balances, people with 3 or 4 products, and inactive members. Let's focus our efforts on these groups first!
- **Wake Up Inactive Members!**: If someone hasn't been active, let's reach out! Maybe a special offer, or just a friendly call to see how we can help them out.
- **Why So Many Products, Yet Leaving?**: This is a head-scratcher: customers with more products seem to churn more. We really need to dig into this. Are they having trouble with multiple products? Is it about fees? We need answers!
- **Germany-Specific Strategies**: Since our German customers are more likely to churn, we might need some special plans just for them. Maybe it's about local competition or unique needs there.
- **Help with Credit Scores**: For customers with lower credit scores, offering financial advice or specific products could really help them improve their situation and feel more loyal to us.
- **Get Personal with Offers**: Let's use what our model tells us to make offers and messages super personal. When customers feel understood, they're more likely to stick around.
- **Watch for Red Flags**: Keep a close eye on big changes in someone's account balance or how active they are. If something shifts, it could be a sign they're thinking of leaving. A quick, proactive check-in could make all the difference!