Project Part 1: Data Exploration and Prediction

0. AI tools -Already filled out the form to report the AI tools used in the project

1. Registration & Staff Meeting - We acknowledge we have submitted the registration form

2. Dataset Selection

Customer churn analysis represents a perfect intersection of business value and data science complexity, where even small improvements in prediction accuracy can translate into significant financial savings. Thus, for this project, we have selected a banking customer dataset that offers rich insights into customer behaviors and characteristics in the bank industry. The dataset captures various aspects of banking customers, including

- Customer's demographic information: Gender, Age, Geography
- Customer's financial metrics: CreditScore, Balance, EstimatedSalary
- Customer's product usage: NumOfProducts, HasCrCard, Tenure
- Customer status: IsActiveMember, Exited

With 8000 observations and 11 meaningful columns (excluding the RowNumber identifier, Surname, CustomerId). We got this dataset from kaggle (public dataset, with permission to use): https://www.kaggle.com/datasets/adammaus/predicting-churn-for-bank-customers

The data structure satisfies all required criteria, featuring a diverse mix of both numeric variables (such as CreditScore, Age, Balance, and EstimatedSalary) and categorical variables (including Geography, Gender, and product usage indicators). This variety enables both a regression task, predicting a customer's credit score (continuous outcome variable), and classification task, predicting customer churn through the binary outcome variable 'Exited'. Importantly, while the dataset includes a 'Tenure' variable, it represents a static measure of customer longevity rather than temporal sequences, ensuring this is not a time-series dataset. Lastly, this raw dataset does not contain any missing value. Thus, we confirmed it meets all criteria from (i) - (v).

3. Holdout Test Set - Acknowledged: 20% of the dataset has been randomly reserved as test set.

4. Choosing outcomes

We plan to use CreditScore as the continuous outcome variable for the regression task. This variable ranges from 376 to 850 in our dataset, providing a meaningful continuous scale that reflects a customer's creditworthiness. Credit scoring is a fundamental metric in banking that is systematically calculated based on various customer attributes and behaviors, making it an ideal continuous variable for predictive modeling. As for the classification problem, we will use the 'Exited' column as the binary outcome variable, where 1 indicates customer has churned (left the bank) and 0 indicates they remain active with the bank. Churn is a well-defined event in banking that directly impacts business performance, making it an appropriate choice for classification modeling. The variable is naturally binary and doesn't require artificial thresholding.

5. Data description and exploration

Studying the relationship between covariates (e.g., Age, Balance, NumOfProducts) and both outcomes, credit scores and customer churn status, can provide valuable insights for banks to optimize their risk assessment and customer retention strategies. Understanding these

relationships can help financial institutions better tailor their products and services to customer needs while managing risk effectively. For example:

- Risk Assessment: Models predicting credit scores (376-850) can help identify which customer characteristics and behaviors contribute to higher creditworthiness, enabling more accurate risk assessment and personalized financial product offerings.
- Customer Retention: Understanding factors influencing customer churn allows banks to identify at-risk customers before they leave, facilitating proactive retention efforts and targeted interventions.
- Decision Making: Models could inform decisions like when to offer specific banking products, which customers might need additional support, or what service improvements could reduce churn rates.

The data was collected from an anonymized banking institution's customer relationship management (CRM) system and made available on Kaggle. The dataset covers customer information across three European countries (France, Spain, and Germany). The sampling strategy was based on existing customer records, which might introduce:

- Sampling Bias: The data may not represent global banking patterns, as it's limited to three European countries. The gender distribution and customer segments included might not reflect the actual bank customer base.
- Selection Effects: Being a public dataset on Kaggle, there could be bias in how the data was chosen and processed for public release. The anonymization process might have affected the representativeness of the data.
- Data Accuracy: While CreditScores (376-850) align with standard ranges and binary variables (HasCrCard, IsActiveMember, Exited) are clearly defined, some values like zero balances could either represent genuine zero balances or be placeholders for missing data, requiring verification in a real-world context.

Our analysis focused on one bank customer dataset of 8,000 records with 11 columns after data preprocessing. Initial feature engineering included encoding Gender (Male: 0, Female: 1), and creating dummy variables for the 'Geography' column, where we dropped the first category in the dummy variable columns to avoid multicollinearity. Then we detected data points whose z-score are outside the specified range (3 standard deviation). By applying Winsorization, we capped extreme values (outliers in ‘CreditScore’, ‘Age’, and ‘NumOfProducts’ columns) below the 5th percentile and above the 95th percentile.

As we plotted the correlation matrix which shows the bivariate relationship between variables, none of covariates showed strong correlation (Pearson correlation coefficient < 0.7) with our two outcome variables (‘CreditScore’, ‘Exited’). Also, those covariates did not show any strong correlation with each other. This does not necessarily mean that these covariates are unimportant for predicting the outcomes. It could imply that the relationships between covariates and outcomes are non-linear or interact in more complex ways than simple linear correlation can capture. Given the lack of strong linear relationships, we may want to consider modeling strategies that can capture complex, non-linear relationships.

In our dataset, subgroups defined by Gender, IsActiveMember, and Geography show notable quantitative differences. For Gender, males have a slightly higher average balance (77,205) compared to females (75,609), but females have a slightly higher average CreditScore (652.54 vs 649.58). For IsActiveMember, inactive users have a higher average balance (76,656 vs 76,309) but a lower average CreditScore (649.46 vs 652.31) compared to active users. Moreover, inactive users have a lower average age than that of active users (37.93 years vs 39.66 years). Geography also reveals distinctions: users in Spain have a lower average balance (61,572), while those in Germany have a much higher balance (119,465). Age differences are minor across geographies, but users in Germany are slightly older on average (39.43 years). These group differences suggest that factors like gender, geography, and membership status will be important for understanding customer behaviors and predicting outcomes.

We also visualized boxplots for Age by Exited and found that older customers are also more likely to exit, as seen by the higher median age of those who left. From the bar plot we can see that female customers may be more likely to stay with the bank compared to male customers, as the count of female customers stay with the bank is higher than that of male customers, and the count of female customers leave the bank (Exited = 1) is lower than that of male customers.

There is no missing value, eg. NULL, NA, or blank values in this dataset.

6. Prediction: Regression

We use RMSE to measure the prediction error, as it is sensitive to outliers and penalizes larger errors more heavily, making it effective for highlighting significant deviations, while also being easy to interpret since it uses the same magnitude/unit as the target variable.

We chose Ordinary Least Squares (OLS) regression as our primary baseline because it is the simplest and most interpretable linear model that uses all available features, providing a fundamental benchmark that requires no hyperparameter tuning and achieved an RMSE of 1.00234 on standardized data.

The remaining 80% were split randomly into 75% training set (60% of total) and 25% estimate test set (20% of total). Then, we used 5-fold cross-validation on the 60% training set for model training, hyperparameter tuning with randomized search as well as model selection. The best performance model selected was then implemented on the 20% estimate test set to get the unbiased estimate of the test error on the 20% hold-out test set which was set aside at the beginning of the project.

Besides the baseline model (OLS), we have also used OLS without Age, Principal Component Regression, Lasso regression, Ridge regression, and Decision Tree as the other 5 modeling strategies. According to the previous feature engineering part, we find that the VIF score of Age is larger than 10 (11.003539), so we can remove Age for reducing Multicollinearity, which may cause unstable and unreliable coefficient estimates, making it difficult to interpret individual variable contributions, and can reduce the precision and generalizability of the regression model. The OLS without Age achieved an RMSE of 1.00183. Principal Component Regression achieved an RMSE of 1.00027 with one component, offering a dimension-reduced perspective to help us understand if simpler representations of our data could be effective. Lasso regression, achieving an RMSE of 1.00015 performs feature selection through L1 regularization and helps prevent

overfitting by shrinking some coefficients to zero, while Ridge regression, with an RMSE of 1.00234, handles multicollinearity well through L2 regularization and provides stability when features are correlated, though it keeps all features in the model. The Decision Tree regressor, achieving an RMSE of 1.01061, offers the advantage of capturing non-linear relationships and feature interactions while being highly interpretable through its tree structure, though it may be prone to overfitting if not properly tuned.

As for Ridge and Lasso regression, we deployed RandomizedSearchCV on both with 20 iterations, with an optimal alpha of 0.0376 for Lasso, and optimal Ridge alpha 5.59.

The **Lasso Regression** emerged as the **Best Performing Model** with RMSE of 1.00015, which outperformed both the OLS baseline (RMSE: 1.00234) and PCR (RMSE: 1.00027). While the improvements in RMSE might appear small since we're working with standardized data, the Lasso model's superior performance combined with its built-in feature selection capability makes it the most attractive choice for our credit score prediction task.

The Decision Tree model didn't perform well (RMSE 1.01061) compared to Lasso (Lasso's 1.00015) was primarily due to higher variance, as evidenced by its complex parameter space (max_depth: 8-20, max_leaf_nodes: 10-100) that likely led to overfitting. The bias seems less of an issue since the model had sufficient flexibility with deep trees and ability to capture non-linear relationships.

When we fit the best Lasso regression model on the 20% estimate test set, the RMSE is 0.98529. As the data remains unseen when we train, tune and select the model, this result will be an unbiased estimate of test error on the holdout set.

We believe there is no deviation from the plan we described in part c. We did not add more model strategies, go back to tune hyperparameters and select the model again after observing results since doing this would introduce bias to our estimate on the test error.

7. Prediction: Classification

We use AUROC as the evaluation metric, which summarizes model performance across all possible classification thresholds, and takes into account both the true positive rate (sensitivity) and false positive rate, providing a more balanced evaluation of classification performance for the imbalance dataset. In our dataset where the target variable 'Exited' shows an imbalanced distribution with 20.5% positive cases (customers who exited) and 79.5% negative cases (customers who stayed).

Logistic regression is used as the baseline model for our binary classification task since the objective of Logistic regression is to estimate the probability (0-1) that a given input belongs to a particular class by applying a logistic (sigmoid) function to a linear combination of input features. Our Logistic regression model returns the AUROC = 0.69.

The remaining 80% were split randomly into 75% training set (60% of total) and 25% estimate test set (20% of total). Then, we used 5-fold cross-validation with StratifiedKFold to maintain class balance/imbalanced nature during CV on the 60% training set for model training, hyperparameter tuning with randomized search as well as model selection. The best performance

model selected was then implemented on the 20% estimate test set to get the unbiased estimate of the test error on the 20% hold-out test set which was set aside at the beginning of the project.

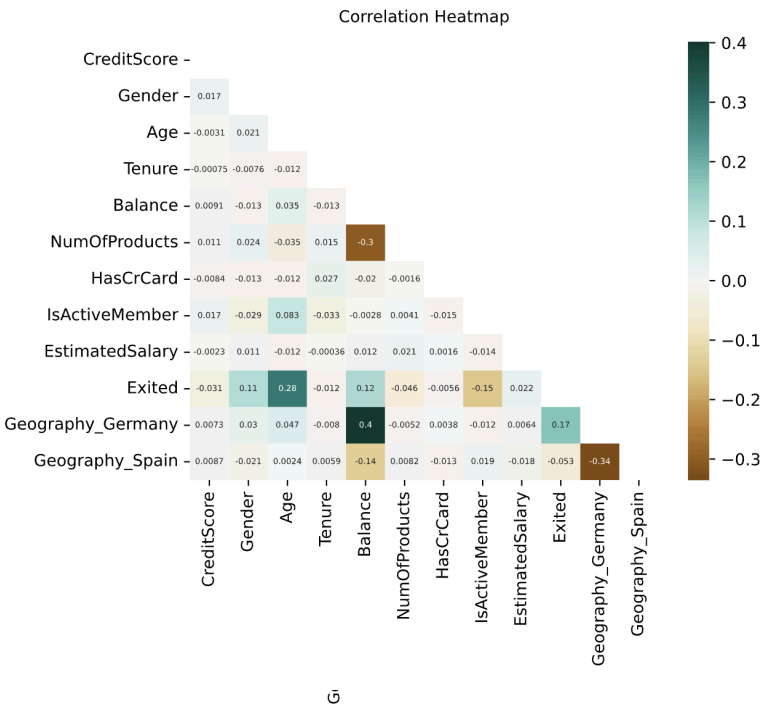
Besides the baseline logistic regression model, we have also used Logistic Regression without Age and CreditScore, Decision Tree, Random Forest, LightGBM as the other 3 modeling strategies. As the VIF of Age and CreditScore are greater than 10 (14.9 and 25 respectively), we drop the two covariates in order to solve multicollinearity issues, achieving AUROC of 0.68. Decision Tree, which achieved an AUROC of 0.78 offers an interpretable approach by recursively partitioning the data based on the most discriminative features, making it easier to understand which variables and thresholds are most important for predicting customer churn. **Random Forest**, achieving an AUROC of 0.82, which is the **Best Model** offers improved prediction accuracy through ensemble learning by averaging predictions from multiple decision trees, while reducing overfitting through bagging and random feature selection at each split, while LightGBM achieved an AUROC of 0.81, offers efficient gradient boosting capabilities through its leaf-wise growth strategy and histogram-based feature discretization, making it particularly effective for handling large datasets with high dimensional features.

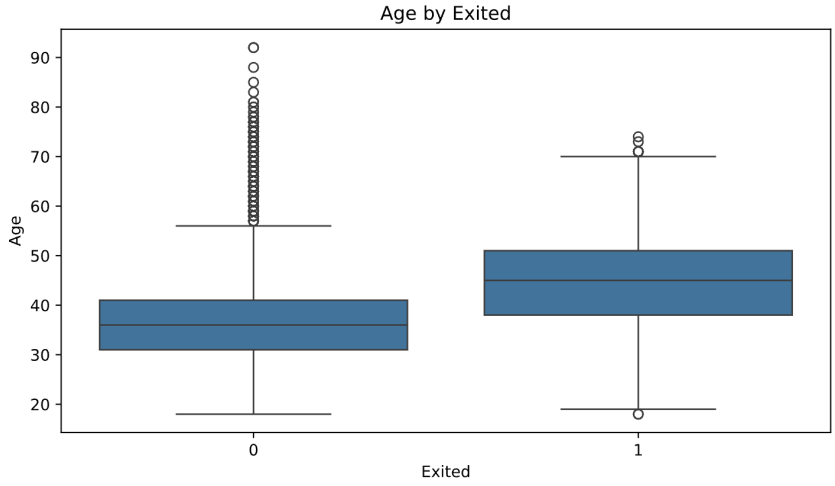
The Decision Tree model didn't perform well (AUROC = 0.78) compared to Random Forest (AUROC = 0.82) was primarily due to higher variance, as single trees are prone to overfitting by growing deep and creating very specific paths through the data. Random Forest addresses this high variance issue by averaging multiple trees using bagging, and employing random feature selection at splits, while maintaining low bias since tree-based models can capture non-linear relationships effectively.

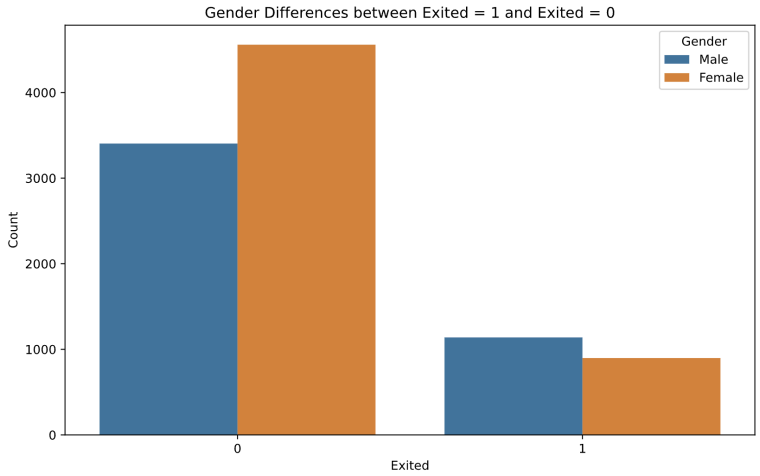
When we fit the best Random Forest model on the 20% estimate test set, the AUROC is 0.82. As the data remains unseen when we train, tune and select the model, this result will be an unbiased estimate of test error on the holdout set.

We believe there is no deviation from the plan we described in part c. We did not add more model strategies, go back to tune hyperparameters and select the model again after observing results since doing this would introduce bias to our estimate on the test error.

Link for code and dataset:
https://drive.google.com/drive/folders/1Nia3tp8cZZaEzEDlKfw0Benw3e9_XoNp


Correlation Heatmap
CreditScore
Gender
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
Exited
Geography_Germany
Geography_Spain
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3


Age by Exited
Age
Exited
0
1
90
80
70
60
50
40
30
20


Gender Differences between Exited = 1 and Exited = 0
Gender
Male
Female
Count
Exited
0
1
4000
3000
2000
1000
0