Bank Churn Prediction Proposal

Project Proposal: Bank Customer Churn Prediction
For this project, I would like to analyze a dataset related to customer churn in the banking industry. The dataset includes various customer characteristics such as age, gender, geography, credit score and balance.

I found this dataset on Kaggle at:
https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn?resource=download.

I have downloaded the data and loaded it into bank_customer_churn.csv, which will be attached to this email.

This dataset contains information on 10,000 customers from a bank, divided into 18 different variables. The variables include a mix of categorical and continuous types. Our dependent variable is whether a customer has churned, indicated by a binary value (1 = churn, 0 = not churn). I plan on segmenting the data into 2 separate parts - one to train the model, and the other to verify predictions from our model. This will allow us to gauge the accuracy of our model.

DIDA Framework for the Case

Data: Daily customer data across multiple bank branches, including features such as credit score, balance, age, gender, and geography.

Insights: What are the main drivers of customer churn (e.g., low credit score, high balance, tenure)?

Decision: Can I predict customer churn based on historical patterns and customer characteristics?

Advantage: Improved customer retention strategies and targeted marketing efforts, reducing churn rate and increasing customer lifetime value.

What is the type of insights in your DIDA?

The insights in our DIDA will focus on probability. I aim to predict the probability of a customer churning based on various demographic and account-related variables.

What is the dependent variable?

The dependent variable I want to predict is churn, which indicates whether or not a customer has left the bank (1 = churn, 0 = not churn).

Is the dataset individual-level data or aggregated-level data?

The dataset is individual-level data, as each row corresponds to a unique customer with associated characteristics.

Do you have the historical values of the dependent variable?

The dataset provides historical information on whether each customer has churned or not, which allows us to build predictive models based on historical churn rates.

Does the dataset have the corresponding binary dependent variable for predicting probabilities?

The dataset does contain a binary dependent variable (churn) that indicates whether the event of interest (customer churn) has occurred.

For the dependent variable, do I have both the cases where the event of interest did happen and the ones where it did NOT happen?

The dataset includes both cases: customers who have churned and customers who have not churned, allowing for a balanced analysis of the factors influencing churn.

Do I have relevant predictors in the dataset? Are they exante as Ill?

Yes, the chosen dataset includes several relevant predictors such as credit score, balance, gender, age, and tenure. These variables can be used to analyze their impact on customer churn.

Does the dataset satisfy the portrait-shape requirement?

The dataset satisfies the portrait-shape requirement with 10,000 observations (rows) and 18 variables (columns), providing a sufficient amount of data for building a reliable predictive model.

Exploration

I did some preliminary exploration of the dataset and identified that it contains a complete set of data with no significant missing values. The dataset is clean and ready for model building. I am confident that the dataset offers valuable insights that will aid in accurately predicting customer churn, allowing us to develop strategies to mitigate it.