Capstone Project 1: Milestone Report

Federico Di Martino

Predicting whether or not a bank customer will churn

Problem Statement

The churn rate is a key metric for almost any business, one that should ideally be minimised. In this specific case, the business will be a bank. The hypothetical client for my work would be that same bank. I will be building a predictor to calculate the probability for an individual customer of the bank to churn. For clarity, I will always use client to refer to the bank and customer to refer to an individual customer of the bank.

Dataset

I will use the Churn Modelling dataset from kaggle. A preliminary inspection reveals that maybe 10 of the variables should be used to build a predictor.

Problem

The problem to solve will be building a predictive model that gives the odds of churning for each individual customer.

Justification

Two possible uses for the ability to accurately predict whether a customer will churn are:

- The client can decide to allocate resources and/or manpower to stop a likely-to-churn customer (or more generally, a group of customers sharing similar characteristics) by identifying any specific issues they have and if necessary altering the business to accommodate their needs.

- The client can triage customers by their churn probability and perhaps decide that for customers are that extremely likely to churn in any case, no further resources need to be allocated to them. Those resources could then be allocated to more salvageable customers.

Data Wrangling

- The dataset consists of a comma separated variable format file that contains 10000 different records and 14 features initially. I have included the codebook at the end of the report.

- I made sure the data was all of the same type within each varialbe column, checked for missing values and checked and also checked for any non-binary values in binary feature columns.

- There are no missing values, however if there had been I would have, after feature selection, discarded any row containing a missing values. I prefer not to perform imputation as that carries some assumptions about the data and I prefer to minimise those. If I had found non-binary values in a binary feature I would have assigned them as missing value.

- I checked the non-binary features for outliers. There were a few, specially under the ‘Age’ category (a few very old customers), but I decided to leave them in as I don’t think there was any justification for removing them. To check the features I generated a plot of each variable paired off with the others. The pair plots are appended at the end of the report.

Data Story and Statistical Data Analysis

- The number of customers which churned out was 2037, or 20.37%, of the total. Finding some hints as to why and to predict if a given customer will churn is the purpose of this project.

- A quick way to get a broad overview of the data is to calculate the Pearson correlation coefficient for each variable pair.

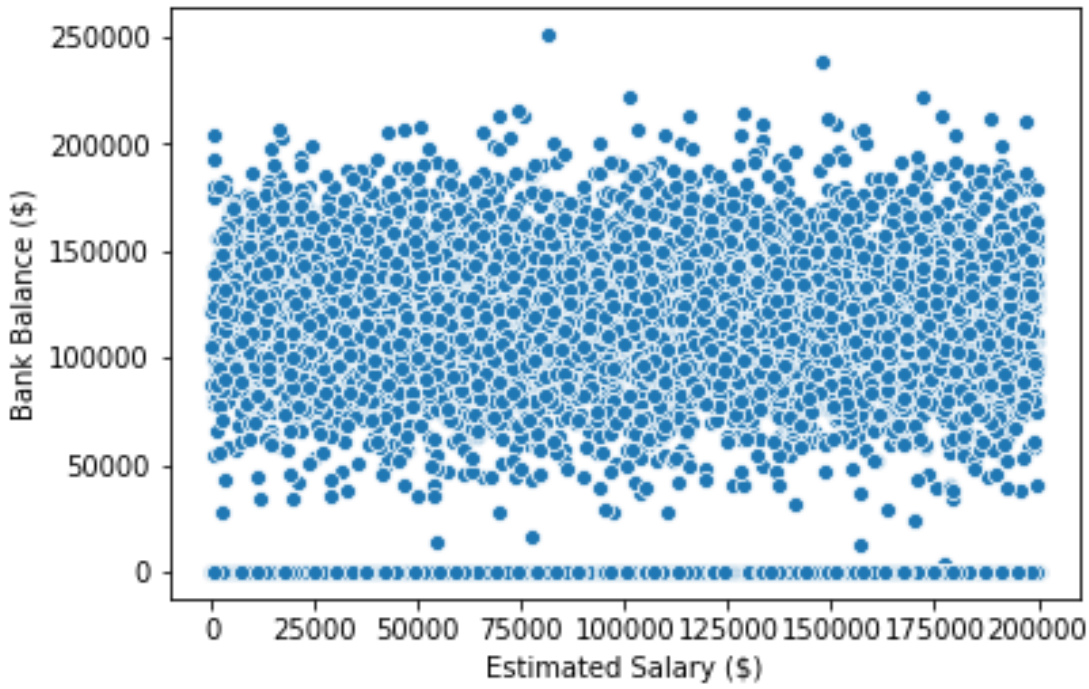
Figure 1: Pearson correlation table.

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	IsFrance	IsSpain	IsGermany
CreditScore	1	0.0029	-0.004	0.00084	0.0063	0.012	-0.0055	0.026	-0.0014	-0.027	-0.0089	0.0048	0.0055
Gender	0.0029	1	0.028	-0.015	-0.012	0.022	-0.0058	-0.023	0.0081	0.11	-0.0068	-0.017	0.025
Age	-0.004	0.028	1	-0.01	0.028	-0.031	-0.012	0.085	-0.0072	0.29	-0.039	-0.0017	0.047
Tenure	0.00084	-0.015	-0.01	1	-0.012	0.013	0.023	-0.028	0.0078	-0.014	-0.0028	0.0039	-0.00057
Balance	0.0063	-0.012	0.028	-0.012	1	-0.3	-0.015	-0.01	0.013	0.12	-0.23	-0.13	0.4
NumOfProducts	0.012	0.022	-0.031	0.013	-0.3	1	0.0032	0.0096	0.014	-0.048	0.0012	0.009	-0.01
HasCrCard	-0.0055	-0.0058	-0.012	0.023	-0.015	0.0032	1	-0.012	-0.0099	-0.0071	0.0025	-0.013	0.011
IsActiveMember	0.026	-0.023	0.085	-0.028	-0.01	0.0096	-0.012	1	-0.011	-0.16	0.0033	0.017	-0.02
EstimatedSalary	-0.0014	0.0081	-0.0072	0.0078	0.013	0.014	-0.0099	-0.011	1	0.012	-0.0033	-0.0065	0.01
Exited	-0.027	0.11	0.29	-0.014	0.12	-0.048	-0.0071	-0.16	0.012	1	-0.1	-0.053	0.17
IsFrance	-0.0089	-0.0068	-0.039	-0.0028	-0.23	0.0012	0.0025	0.0033	-0.0033	-0.1	1	-0.58	-0.58
IsSpain	0.0048	-0.017	-0.0017	0.0039	-0.13	0.009	-0.013	0.017	-0.0065	-0.053	-0.58	1	-0.33
IsGermany	0.0055	0.025	0.047	-0.00057	0.4	-0.01	0.011	-0.02	0.01	0.17	-0.58	-0.33	1

- At first glance there aren't many strong correlations. The most important row to observe is how Exited correlates with the others, as that is what we want to be able to later predict. An important relationship to examine further is that of Exited (churn) with Age.

- Another thing to investigate is how Balance and Estimated Salary are almost entirely uncorrelated! This seems very unintuitive and merits a further in depth look, although it may be tangential to the main objective of predicting churn.

Figure 2: Bank Balance against Estimated Salary



- The project aims to predict the likelihood of a customer churning (Exited =1 in the data). Hence the particularly significant variables are the ones that correlate the most with Exited. Although there are no extremely strong correlations, the Gender, Age, Balance, IsActiveMember and IsGermany variables are the ones that show the most correlation and should be investigated the most.

- I think the narrative that emerges here is one of lack of clarity. On first inspection, whether a customer churns out is weakly correlated to anything else. I think a compelling story can be made out of the need to dig deeper. It would have been interesting to build time series data, had date/time data been available.

- The next steps will be constructing machine learning models to predict customer churn.

Codebook

RowNumber: Row Numbers from 1 to 10000. Variable dropped when data loaded into Python environment because it is redundant.
CustomerId: Unique IDs for bank customer identification.
Surname: Customer's surname.
CreditScore: Credit score of the customer.
Geography: The country from which the customer belongs.
Gender: Male or Female.
Age: Age of the customer.
Tenure: Number of years for which the customer has been with the bank.
Balance: Bank balance of the customer.
NumOfProducts: Number of bank products the customer is utilising.
HasCrCard: Binary Flag for whether the customer holds a credit card with the bank or not.
IsActiveMember: Binary Flag for whether the customer is an active member with the bank or not.
EstimatedSalary: Estimated salary of the customer in Dollars.
Exited: Binary flag 1 if the customer closed account with bank (churned) and 0 if the customer is retained.

Pair Plots

Figure 3: Pair plots of each variable against each other. When a variable is plotted against itself a histogram is generated.