Predicting European Bank Churn

Jared Diehl

February 25, 2021

Contents

1 Introduction

Churn, short for churn rate, is the rate at which customers stop doing business with an entity. [2] There can be several factors that affect churn, including age, gender, and salary. Europe houses some of the world's largest economies, such as France, Germany, and Spain. A business may want to predict when their customers might leave, especially a bank. This project should help make these predictions.

Motivation This project aims to develop a model using supervised learning, to predict which customers are more likely to leave their bank.

2 Data

2.1 Description

To explore why customers leave their bank, I have chosen a churn modelling dataset from Kaggle relating to banks. This dataset is made for binary classification, and it is tabular. The author did not state how or when the data was collected.

Source: `https://www.kaggle.com/shrutimechlearn/churn-modelling`

It consists of 10,000 rows and 14 columns:

- Row Number
- Customer ID
- Surname
- Credit Score
- Geography
- Gender
- Age
- Tenure
- Balance
- Product Count
- Has Credit Card
- Is Active Member
- Estimated Salary
- Exited

2.2 Cleaning

Data cleaning refers to identifying and correcting errors in the dataset that may negatively impact a predictive model. [1] Fortunately, there are no missing values, so all 10,000 rows can be used. For columns, some are irrelevant for helping predict churn, so they will be removed. Some names were changed to provide better meaning and follow the convention of lowercase words separated by underscores.

The resulting columns are the following:

- credit_score
- geography
- gender
- age
- tenure
- balance
- product_count
- has_credit_card
- active
- estimated_salary
- exited

2.3 Visualization

Data visualization is the representation of data in a graph, chart, or other visual formats. It communicates the data with images. This is important because it allows trends and patterns to be more easily seen. [3] In Figure 1, a correlogram is used to visualize churning amongst the input features. This way, we can see which are correlated with each other.

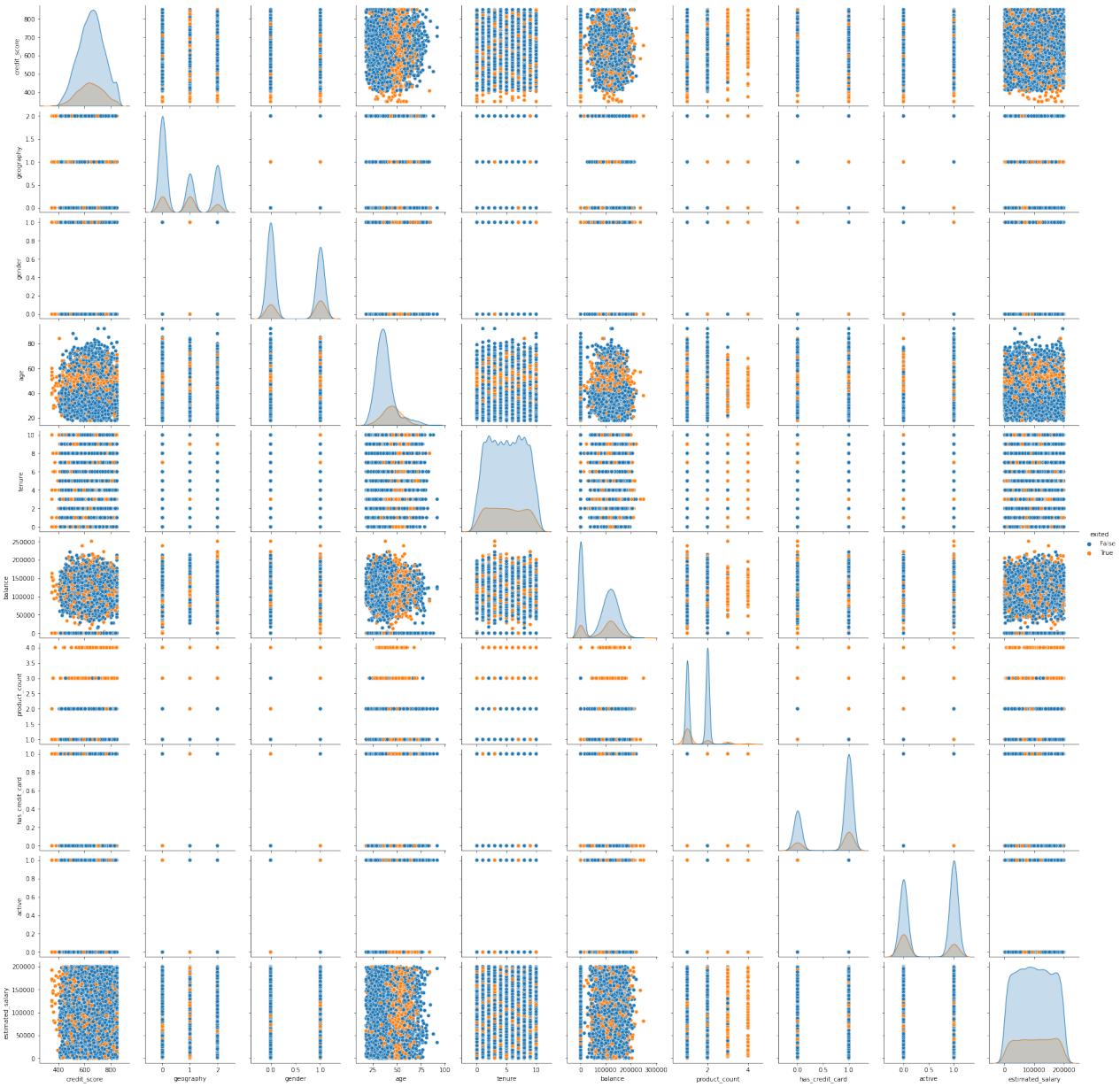

Figure 1: Feature Correlations

2.3.1 Input

Input refers to the features that are tuned during training in the model. For this project, there are ten input features. In Figure 3, we see how they are distributed. For `credit_score`, `age`, and `balance`, we also see that they are normally distributed.

	credit_score	geography	gender	age	tenure	balance	product_count	has_credit_card	active	estimated_salary	exited
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	650.528800	0.746300	0.454300	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	96.653299	0.827529	0.497932	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	350.000000	0.000000	0.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	584.000000	0.000000	0.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	652.000000	0.000000	0.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	718.000000	1.000000	1.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	850.000000	2.000000	1.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Figure 2: Feature Statistics

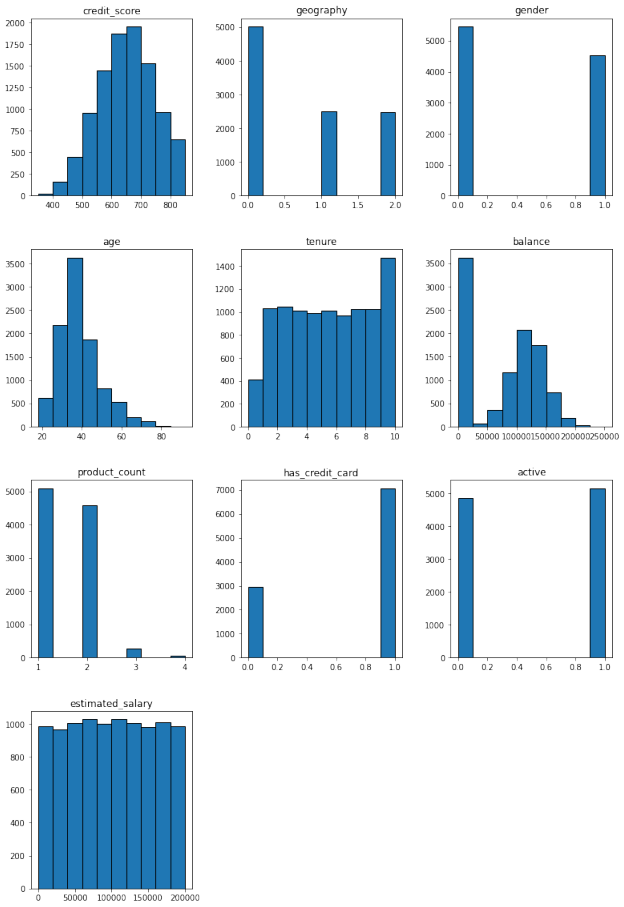

Figure 3: Unnormalized Input Distributions

2.3.2 Output

Output refers to the feature that will be predicted given some input. For this project, there is one output label. The `exited` output label is churn. As we can see in Figure 4, there is considerably less churning (20.37%) than no churning (79.63%). Hence, the data is moderately imbalanced. However, it is acceptable for now.

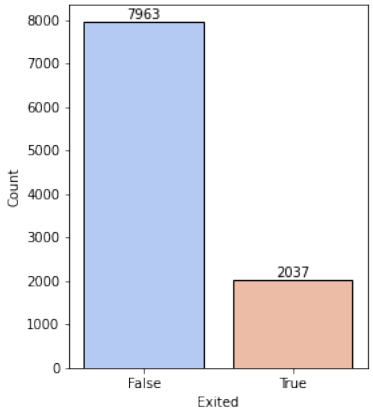

Figure 4: Output Distribution

2.4 Splitting

Splitting the data into two sets will allow the model to learn better. The dataset is randomly shuffled and split into those sets: training (80%) and validation (20%). The reason for shuffling is to help decrease the chance of biases during the training process.

2.5 Normalization

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1, which helps the model make better predictions faster. The data will be normalized using the mean normalization technique.

Mean Normalization Formula

$$X_{new} = \frac{X - X_{mean}}{X_{max} - X_{min}}$$

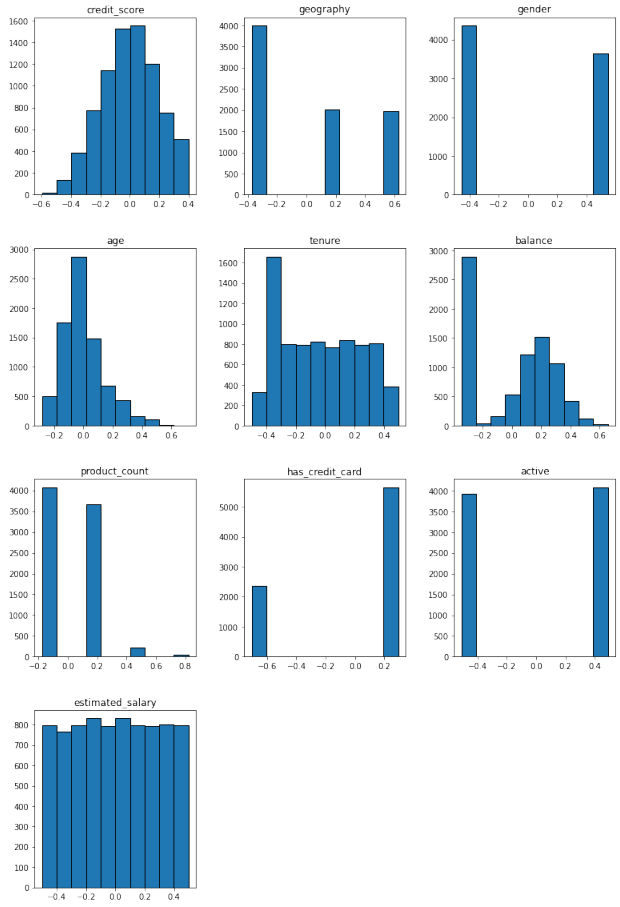

Figure 5: Normalized Input Distributions

To be continued...

References

[1] Jason Brownlee. *How to Perform Data Cleaning for Machine Learning with Python*. URL: https://machinelearningmastery.com/basic-data-cleaning-for-machine-learning/. (accessed: 02.25.2021).

[2] Jake Frankenfield. *Churn Rate*. URL: https://www.investopedia.com/terms/c/churnrate.asp. (accessed: 02.25.2021).

[3] Import.io. *What is Data Visualization and Why Is It Important?* URL: https://www.import.io/post/what-is-data-visualization. (accessed: 02.25.2021).