**IN402 Unit 7 Assignment**

Laurence T. Burden

Purdue Global University

IN402: Modeling and Predictive Analysis

Steward Huang

December 29, 2023

The dataset provided contains details for banking customers with a target variable of whether each customer remained or left, also known as churn (Iyyer, 2019). There are a total of 11 features across 10,000 observations. The target variable is titled “Exited.”

There are eight variables that would work for building a model. These are CreditScore, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, and EstimatedSalary. The descriptive statistics of these variables are listed in table 1.

**Table 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Type | Mean | Range | Standard Deviation | Variance |
| CreditScore | Continuous | 650.53 | 350-850 | 96.65 | 9,341.86 |
| Age | Continuous | 38.92 | 18-92 | 10.49 | 109.99 |
| Tenure | Continuous | 5.01 | 0-10 | 2.89 | 8.36 |
| Balance | Continuous | 76,485.89 | 0-250,898.09 | 62,397.41 | ~3.89x109 |
| NumOfProducts | Discrete | 1.53 | 1-4 | 0.58 | N/A |
| HasCrCard | Discrete | 0.7 | 0-1 | 0.46 | N/A |
| IsActiveMember | Discrete | 0.52 | 0-1 | 0.5 | N/A |
| EstimatedSalary | Continuous | 100,090.24 | 11.58-199,992.48 | 57,510.49 | ~3.3x109 |

Each of the variables listed above would be useful in a predictive model. The high variance seen in Balance and EstimatedSalary could impact the predictive model. The data does not need to be standardized for a logistical regression model. Logistical regression models are not sensitive to variance between variables.

Standardization should be utilized for this dataset before creating a SVM model. This is because the range of data in each variable is so vast. This will allow the model to be more accurate in its predictions by bringing the data into a more uniform set of values.

Two more variables that can be added to the model are Gender and Geography. These are discrete variables with two and three possible values, respectively. Dummy variables should be created before using these for modeling, though. This will allow for a value of only 1 or 0 to be present in the column during the analysis.

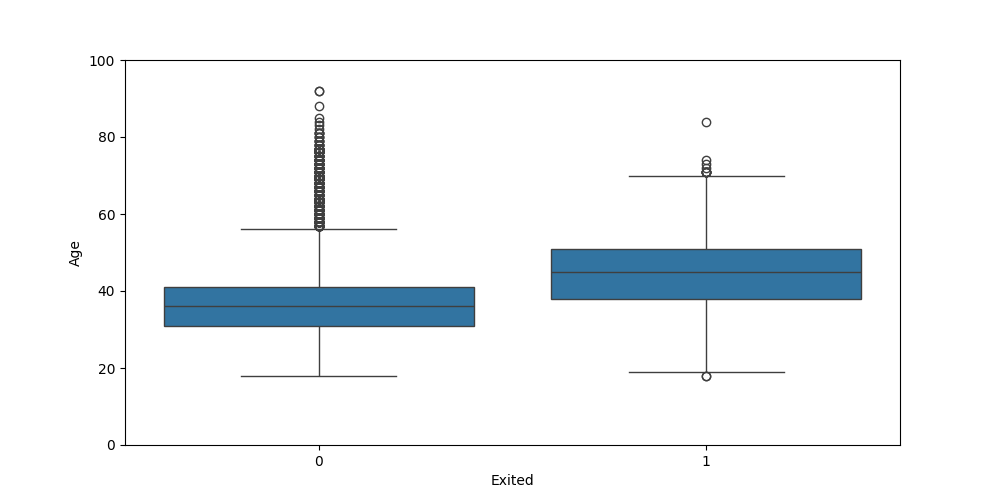
This dataset contains 13 features, but only 10 are relevant to predicting the 1 target variable. There are five continuous variables and three discrete variables that can be described using descriptive statistics in a meaningful way. Two variables would benefit from the creation of dummy variables before creating a model. All visualizations are placed in Appendix A.

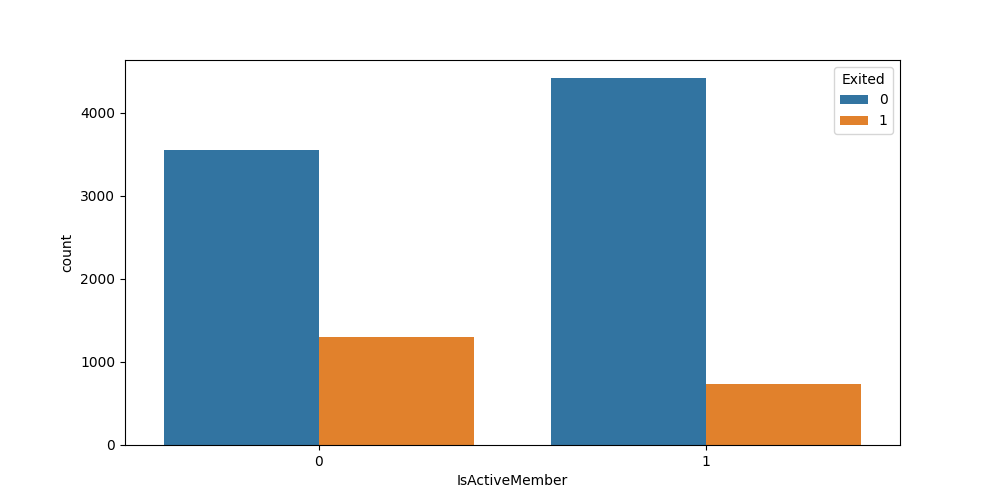
**References**

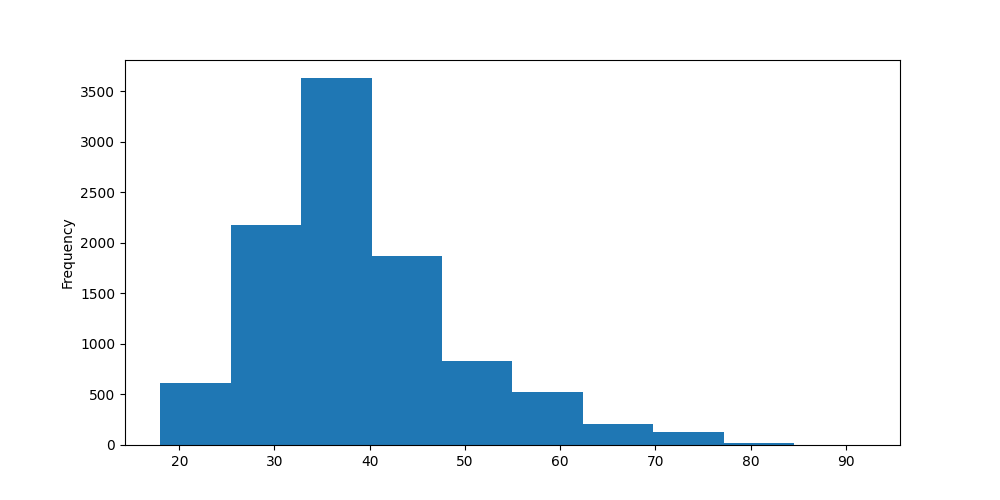
Iyyer, Shruti. (2019). *Churn Modelling* [Data set]. <https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling?resource=download>

**Appendix A**

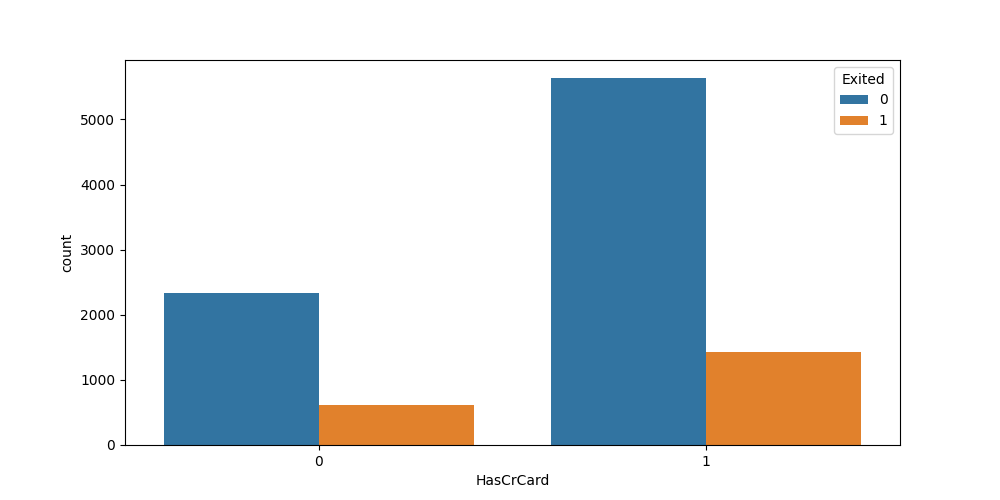
**Figure 1**

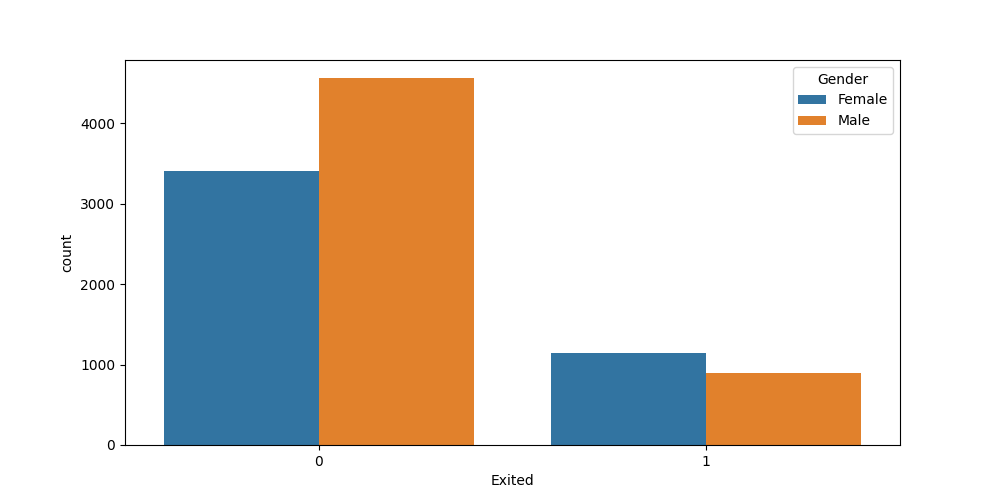
****

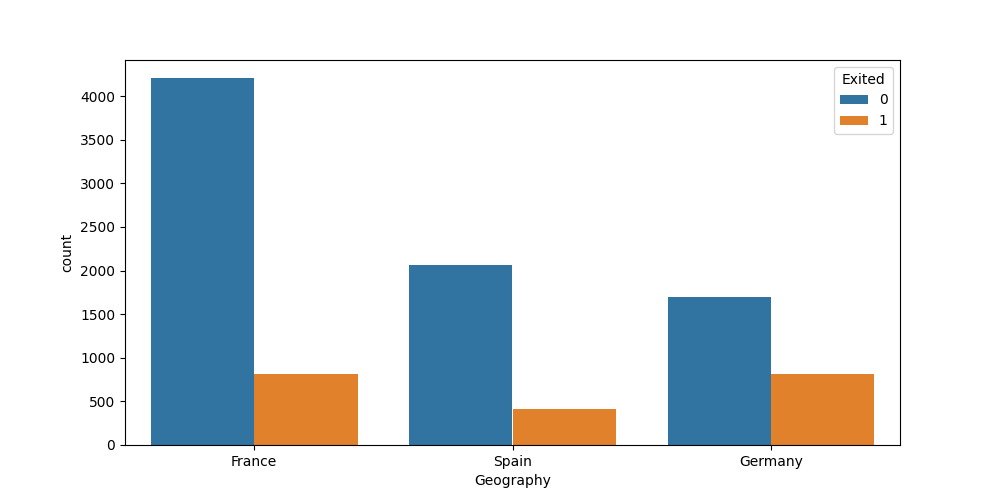
**Figure 2**

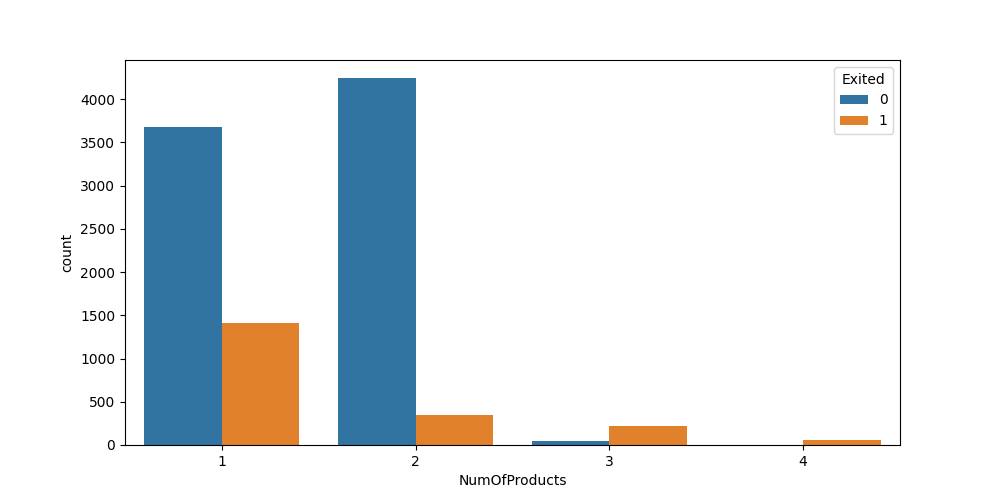
**Figure 3**

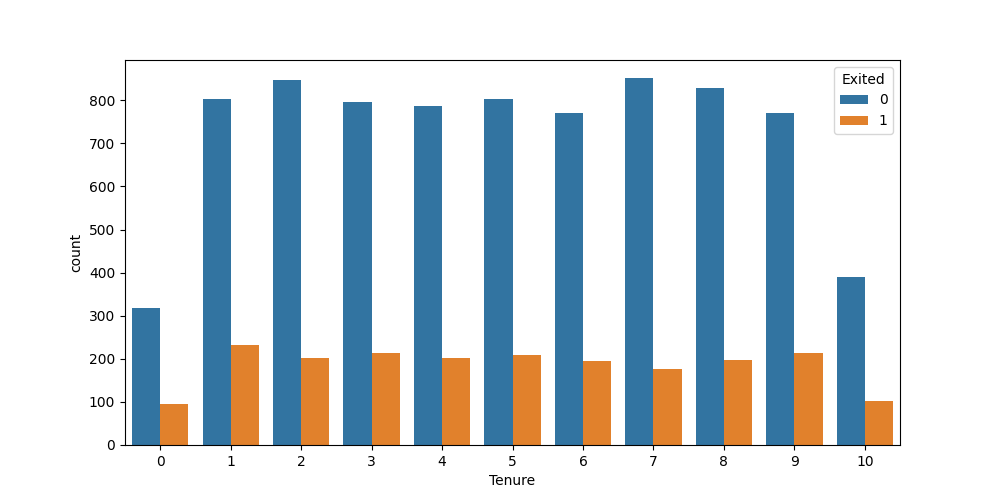
**Figure 4**

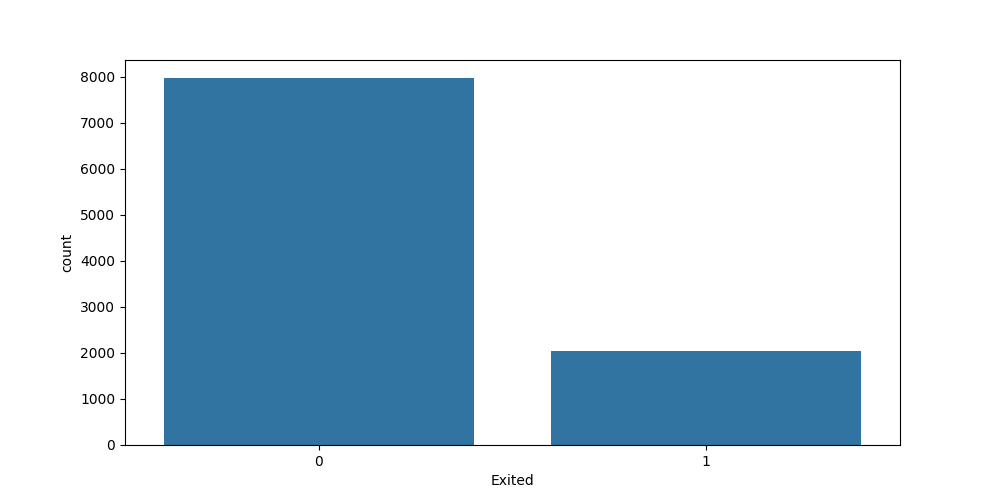
****

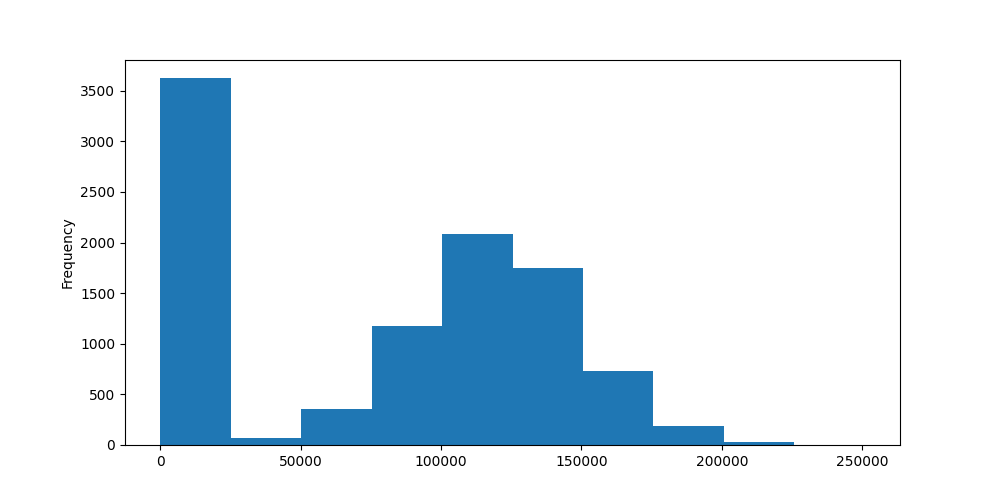
**Figure 5**

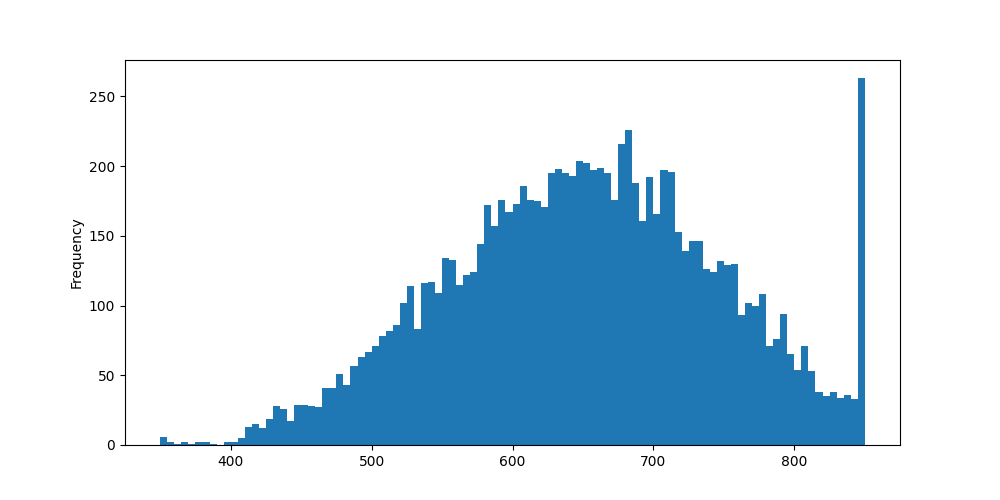
**Figure 6**

**Figure 7**

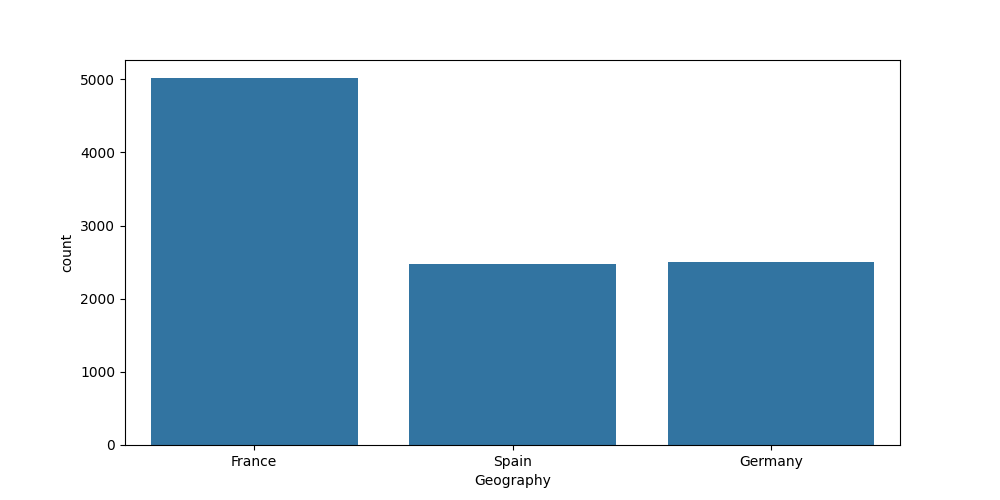
**Figure 8**

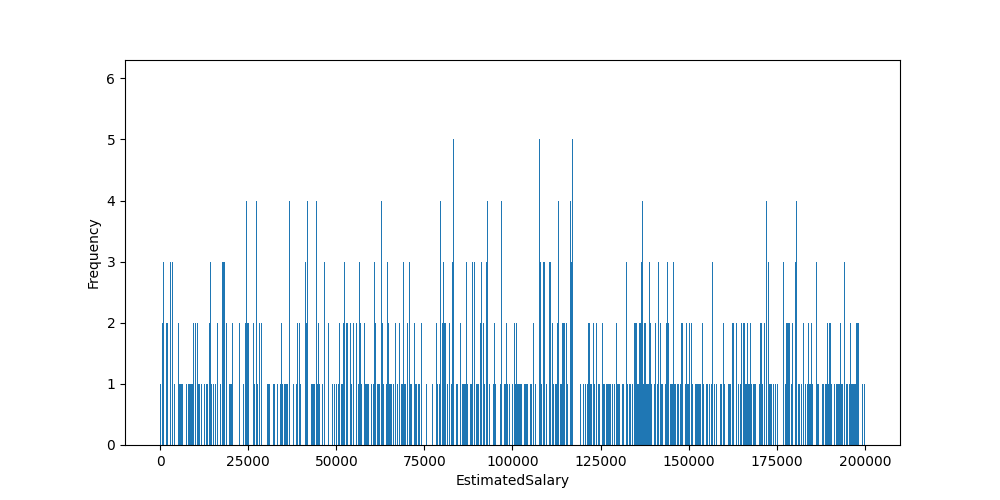
**Figure 9**

**Figure 10**

**Figure 10**

**Figure 11**

****

**Figure 12**