Report: Statistical Inference on Bank Customer Churn dataset

The Dataset is about bank customers churning and can be found on Kaggle:

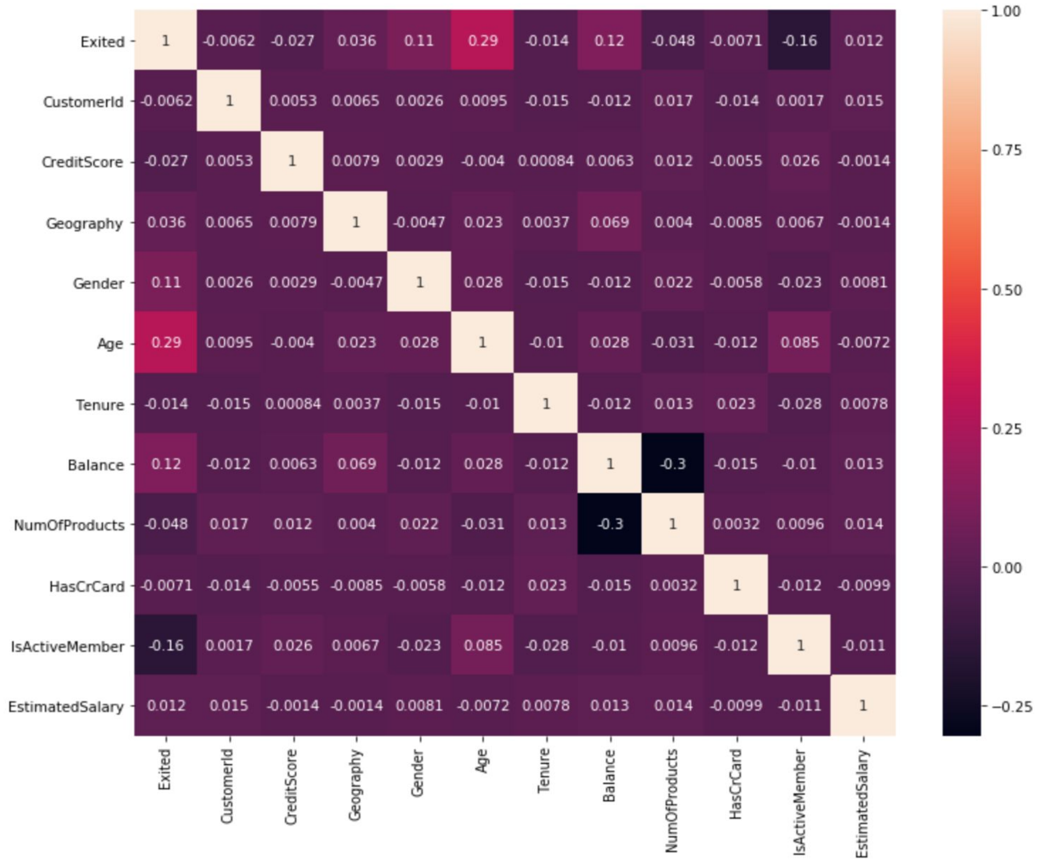
https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling

Disclaimer: The dataset above is simulated.

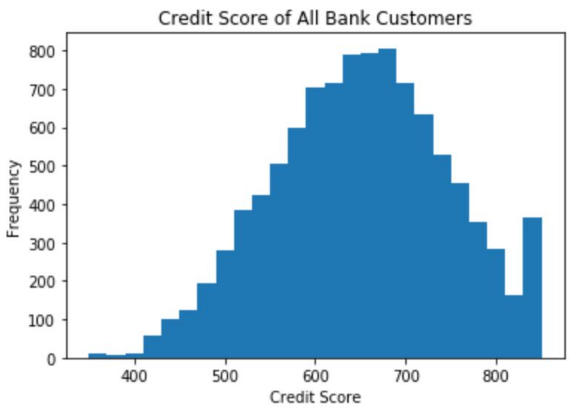
The aim of this exercise was to investigate the data in the dataset further through a statistical lens and perform a test to see if there is any statistical difference in the means of credit score of customers who exited compared to the mean credit score of the entire bank customer population.

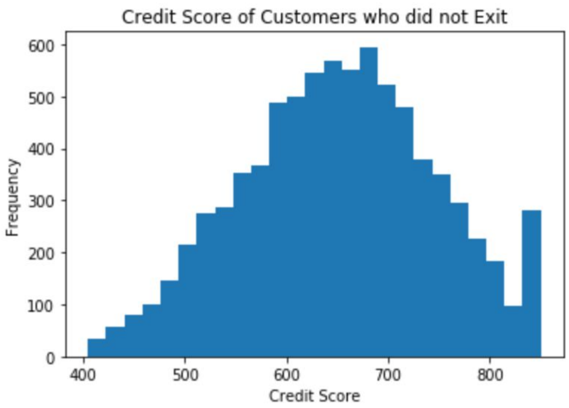
I was curious about the proportion of customers who had stayed in the bank compared to the customers who had exited. I found that 79.6% of customers stayed at the bank while 20.4% of customers exited the bank. The mean credit score of customers who stayed was 651.85 compared to 645.35 for customers who exited the bank.

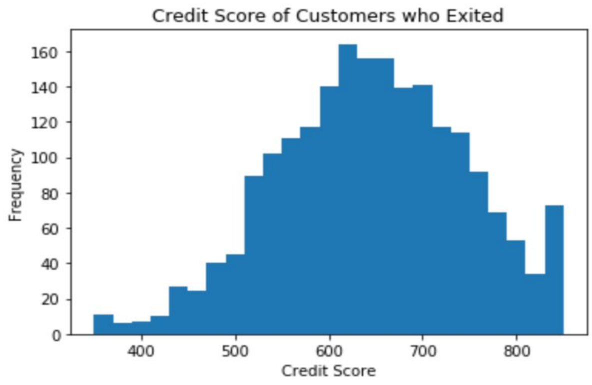
I wanted to explore the variables in the dataset to see if any meaningful correlation existed between them. A correlation value close to -1 indicates strong negative correlation, values close to 0 indicate no/very weak correlation and values close 1 indicate a strong positive correlation between variables. The ‘Exited’ and ‘Age’ variables had the highest correlation value of 0.29, which does not suggest strong correlation between the pairs as the value is close to 0. All the other variables had very weak correlation values as can be seen below in the heatmap.



I also wanted to explore how the credit score distributions looked like for customers who had exited, who had stayed and the entire bank customer population. The credit score distribution of customers who exited appears different than the other two distributions.







Next, I performed a one-sample t-test to test whether a population mean is significantly different from some hypothesized value. In this case, we are going to test to see whether the average credit score of customers that exited differs from the average credit score of the entire bank population. For this test, my Null and Alternate hypothesis were:

Null Hypothesis: The null hypothesis would be that there is no difference in credit score between customers who exited and all bank customers.

Alternate Hypothesis: The alternative hypothesis would be that there is a difference in credit scores between customers who exited and all bank customers.

Before performing the test, I found the mean credit score of all bank customers (650.5288) and the mean credit score of bank customers who exited (645.3514). I conducted the t-test at 95% confidence interval and checked if the null hypothesis (sample comes from the same distribution as the bank customers population) is correctly rejected.

The t-distribution left quartile range was -1.96112925575354 and right quartile range was 1.9611292557535396. The test result showed the test statistic 't' is equal to

-2.329. T is simply the calculated difference represented in units of standard error and tells us how much the sample mean deviates from the null hypothesis. We can reject the null hypothesis if the t-statistic lies outside the quantiles of the t-distribution corresponding to our confidence level and degrees of freedom.

A p-value of 0.019946347165310532 means we'd expect to see data as extreme as our sample due to chance way less than 5% of the time if the null hypothesis was true. In this case, the p-value is lower than our significance level α (equal to 1-conf.level or 0.05) so we should reject the null hypothesis.

Based on the statistical analysis of a one sample t-test, there seems to be some significant difference between the mean credit score of bank customers who exited and the entire bank customer population. The low P-value of 0.019946347165310532 at a 5% confidence interval is a good indicator to reject the null hypothesis.

Although, we should be mindful that this does not necessarily mean that there is practical significance. We can conduct more experiments or maybe collect more data about the bank customers in order to get more accurate insights. I would recommend getting more variables from the bank customers database that could have more impact on determining bank customers exiting and credit score such as satisfaction levels with the bank, APR rates offered by the bank, etc.