Bank Customer Analysis

Ilia Mitrokin, Israel Adewuyi

March 8, 2024

Innopolis University

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Hypothetical Introduction.

In recent years, the banking industry has *hypothetically* witnessed a surge in account closures. Pressed by the need to make informed decision based on data, The Bank of Innopolis, owned by Vladmir Ivanov, who himself is a Giga Chad in the field of data analysis and Machine Learning sought to first of all understand the trend and get a general sense of what's going on. He asked his Senior VP of Customer Success, Vladmir Bazilevich to immediately get on the problem and as young interns in the bank, we were given access to the customer data, every data point that might be related to account closure and tasked to find the underlying factors driving customer churn.

According to [1], who analyzed the same dataset, they found out that the factor most correlated with if customers closed their account or not, was Age. The researchers also listed two other factors, but these will be ignored.

Building upon the foundational work of these researchers, our study seeks to further investigate the relationship between age group and account closure tendencies. To better understand our bank clients, we thought it would be a good idea to also investigate the standard of living of our clients in different countries, to get a sense of their economic well-being.

In summary, our study seeks to investigate the following questions:

- Are there any correlations between the age of customers and if they exit their accounts or not?
- What is the relative standard of living of customers, in the countries they are from?
- Are there any correlations between user's attributes (e.g with tenure and balance)
- Does the CreditScore follows some data distributions (e.g. normal, gamma)?

Chapter 2

Data

The dataset we used was taken from Kaggle. It contains information on bank customers who either left the bank or continue to be a customer. Examples of the attributes:

- **Credit Score:** A numerical value representing the customer's credit score
- **Geography:** The country where the customer resides (France, Spain or Germany)
- **Age:** The customer's age
- **Tenure:** The number of years the customer has been with the bank
- **Balance:** The customer's account balance.

Chapter 3

Theory / Statistical Techniques

3.1 Question 1

To answer the Question 1, we decided to first of all, quickly confirm that age was the most correlating feature of all the non-numerical features, with a correlation matrix.

We proceeded to divide the age into groups. We thought it would be easier to divide the ages into groups of 10. We do note however that we are not certain if this is the most optimal division for the ages, count the number of customers in each age group who exited or did not exit and then organized them into a pandas dataframe.

To test if there is a significant relationship between the age groups and the occurrence of exiting, we used the Chi2-contingency test, available in Scipy.stats package.

The Chi-squared contingency test is a statistical method used to determine whether there is a significant association between two categorical variables, in this case, the age group and the fact of if customers exit their accounts or not. It calculates the difference between the observed frequencies and the expected frequencies under the assumption that the variables are independent. If the difference is large enough, it suggests that the variables are likely not independent, indicating a significant association.

To go further, we decided to test if there is a correlation between the individual age groups and the occurrence of exiting.

3.2 Question 2 - 4

After analysing the data, we decided to use Credit score as a proxy for the standard of living. While there are issues with this association,

In assessing the standard of living across different countries like Germany, France, and Spain, we often utilize available metrics such as credit scores. While credit scores primarily gauge financial solvency, they indirectly reflect the standard of living, as they are crucial in accessing financial services like loans and credit cards. In the absence of direct measures, credit scores serve as proxies for economic well-being and that is what we utilized in this study, to measure the standard of living.

We conducted a t test, to check, as an example, if Germany has a higher credit score and as a proxy, higher standard of living, compared to the two other countries. This is a statistical test used to investigate whether the difference between the response of two groups is statistically significant or not.

Brief description of the T-test (two sample, because we compare 2 samples from the same population): A two-sample t-test investigates whether the means of two independent data samples differ from one another. The null hypothesis is that the means of both groups are the same. Unlike the one sample-test where we test against a known population parameter, the two sample test only involves sample means.

We also made an attempt at examining if there was a correlation between the credit score of customers and their salaries. We did this by taking random samples from the three countries and plotting the customers' salaries against their credit score. Additionally, we tried to plot all attributes to investigate their dependencies.

Finally, we examined if the credit score follows a Normal distribution. For this task we use Kolmogorov-Smirnov test. It answers the question: What is the probability that this collection of samples could have been drawn from that probability distribution? We tried to substitute normal (as the most popular in our world. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. Additionally, exponential and gamma distributions were tested). In the end Kolmogorov-Smirnov test for two samples was used to check that the data are from the same (unknown) data distribution.

Brief description for the algorithm of Kolmogorov-Smirnov test:

- Calculate ECDF from the sample
- Calculate KS Statistic: largest absolute difference between the two distribution functions across all x values
- Use table to check critical values Table

Chapter 4

Statistical Tools

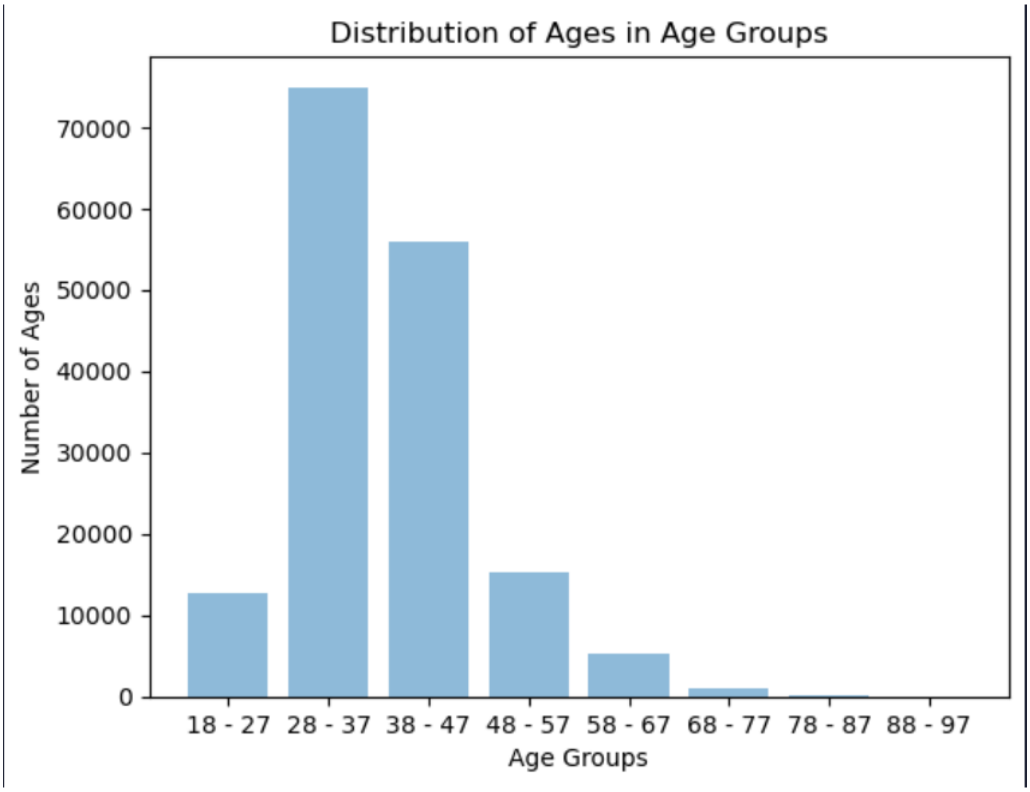
In this study, we used the following tools:

- Matplotlib: plotting
- Scipy: Pefrorm tests
- Pandas: reading and working with a data frame
- Seaborn: creating plots
- Numpy: necessary functions
- ChatGPT, google, labs: search methods, ideas.

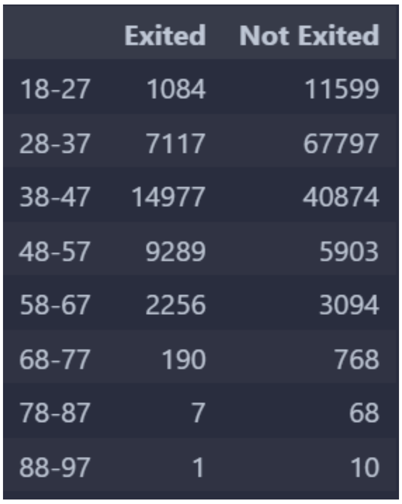
Chapter 5

Results

5.1 Question 1



This represents the distribution of the members of the different age groups.



	Exited	Not Exited
18-27	1084	11599
28-37	7117	67797
38-47	14977	40874
48-57	9289	5903
58-67	2256	3094
68-77	190	768
78-87	7	68
88-97	1	10

This represents the number of people, in each age group who exited or did not exit their accounts with the bank.

For the Chi2 contingency test, we got the following test statistic:

- Test statistic : 24367.955660083328

- p-value : 0.0

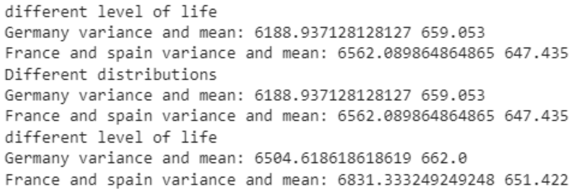
- DoF : 7

This allows us reject the null hypothesis that there is no significant association between age and account closure.

We did not access how the number of divisions to the age group affected the results, so in the future, this might be something to consider. The p-value of 0, while theoretically possible, is practically weird.

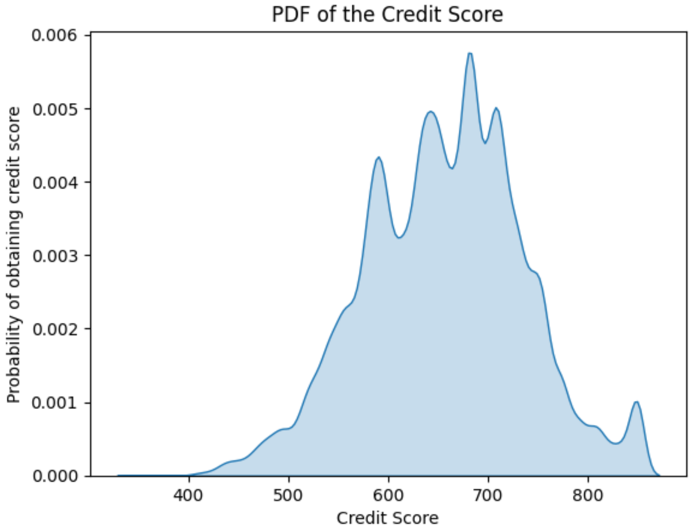
5.2 Question 2

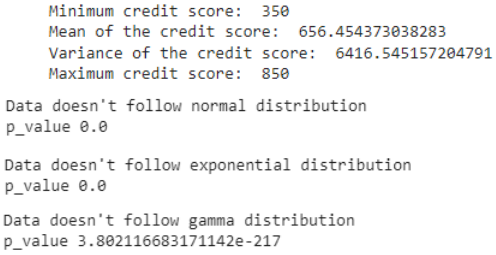
While investigating the standard of living in given countries, we obtain such statistics: 99 of 100 cases shows that the Germany and France with Spain have the same standard of living, while 1 case shows that Germany is better (which is slightly, but true according to the wikipedia). For such cases always Germanies mean metrics is stronger.



```
different level of life
Germany variance and mean: 6188.937128128127 659.053
France and spain variance and mean: 6562.089864864865 647.435
Different distributions
Germany variance and mean: 6188.937128128127 659.053
France and spain variance and mean: 6562.089864864865 647.435
different level of life
Germany variance and mean: 6504.618618618619 662.0
France and spain variance and mean: 6831.333249249248 651.422
```

Next, we tried to define the probability distribution data. Unfortunately, our attempts doesn't provide success: Credit Score does not follow normal, exponential, or gamma distribution. PDF of the CreditScore and results of tests are shown below:

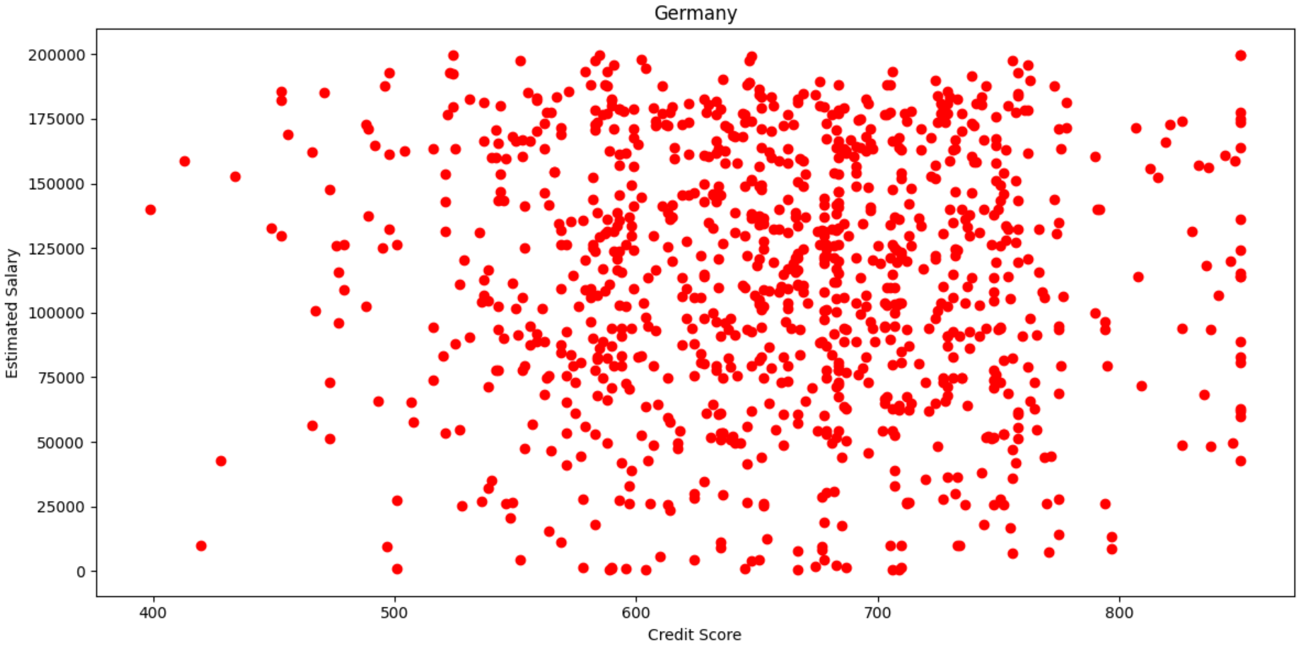




```
    Minimum credit score:  350
    Mean of the credit score:  656.454373038283
    Variance of the credit score:  6416.545157204791
    Maximum credit score:  850

Data doesn't follow normal distribution
p_value 0.0

Data doesn't follow exponential distribution
p_value 0.0

Data doesn't follow gamma distribution
p_value 3.802116683171142e-217
```

The only thing that we can say is that the data are from the same distribution (again, in 99 of 100 cases, because sometimes Germany score is higher).

To explore dependencies, we first plot Credit Score and Estimated salary graph to check our assumption that higher creditScore means higher salary. Unfortunately, this is not true. People with the same credit score can have different salary. The only thing is that for high credit score (>800) minimum salary starts with much bigger value than for average credit score (approximately 656).



Finally, we build all possible correlations and found that this is not significant amount of that for this dataset. There are several observations from this dataset:

- people with low balance on their card seems not to continue using their card in the bank (Balance - Exited graph).
- Current tenure doesn't influence the decision of using card further or not. (Tenure - Exited)
- Non-Active user may continue to have a card in the bank while active user may stop it (IsActive - Extied)
- Balance on the card seems to be more for users with more salary (Balance - EstimatedSalary)

CONCLUSION

Chapter 6

Conclusion

Let's finalize the results and provide conclusions for the result section.

- Are there any correlations between the age of customers and if they exit their accounts or not?

  There is indeed a correlation between the age groups and if customers close their accounts or not.

- What is the relative standard of living of customers, in the countries they are from?

  Customers from Spain, France, and Germany shows that their standard of living is high among other countries from all over the world. Additionally, residents of Germany sometimes shows that their living standards little bit higher than for France and Spain.

- Are there any correlations between user's attributes (e.g with tenure and balance) No, there are no strong correlation between the data attributes. It shows, that every person is individual and has its own properties

- Does the CreditScore follows some data distributions (e.g. normal, gamma)?

  From the tests that we performed we can't draw the correct probability distribution of data. Only thing that we found is that data from the same (unknown) distribution, which sometimes doesn't true. Following reasons may be the explanation for these outcomes:

  Outliers in the data. (especially for the testing that the data follows the same distribution). They can skew the data, making it deviate from the expected distribution.

  Data Transformation. Maybe we should transformed the data and investigate it more times and it will fit a specific distribution.

  Data itself. The dataset for this competition (both train and test) was generated from a deep learning model trained on the Bank Customer Churn Prediction dataset. Feature distributions are close to, but not exactly the same, as the original. We don't know original distribution of the data, maybe it doesn't follow any distribution at all.

Bibliography

[1] RTIEBT: Bank customer churn prediction a comparison between classification and evaluation methods. https://www.diva-portal.org/smash/get/diva2:1435454/FULLTEXT01.pdf (6 2020)

Appendix A

Code