**ABT Case Analysis**

University Canada West

BUSI 652 (Section- 12): Predictive Analysis: What Works?

Prepared by: Aditya Arte

Student ID: 2305505

Date: May 12, 2024

**Table of Contents**

[**Introduction 3**](#_e726fvvywlud)

[**Model Proposal and Implementation 4**](#_oc67p7q17xz7)

[**Practical Analysis 6**](#_6c39rp3m89wv)

[**Conclusion 10**](#_fn49wqd3awha)

[**References 11**](#_tgqs2ojpfgcx)

# **Introduction**

In the real world, professionals are not presented with any specific analytical problem statement; they are provided with a dataset. They need to make sense of the data and identify the potential problem or risk. In this academic paper, the author has worked towards finding a real-world analytical problem, understanding its underlying concepts, proposing a model to find a solution, and conducting a practical analysis (Sghir et al., 2022).

The data available to the author contains a bank's historical data on its customers. The author is expected to predict customer churn. In the business context, customer churn is when customers discontinue using a particular company's services, such as a bank. Churn can result from customers' dissatisfaction with the product or service, better offerings from competitors, changes in personal needs, or discontinuation due to disengagement (Stephen, 2024).

# **Model Proposal and Implementation**

In this case study, the author considers himself an employee of the bank, and with the provided data, he will predict customer churn. This case study or assignment will help the author's employer devise plans or design contingencies to stop customer churn, retain maximum customers, or make more profit.

The author has identified three different prediction models. The most viable model, according to the author, is customer churn. This model will be a binary predictive model, which predicts if a customer will be retained. The necessary features needed for implementing customer churn predictive are transactional data, which includes account activity and transactions; demographic data, which includes age, gender, and geographical location; customer interactional data, which includes customer service calls and complaints; and product usage, which includes number of products availed and for what tenure. The author has data for the rest of the features except for customer interactional data. The author plans to procure the remainder of the data features by requesting the customer service department to provide data from their transactional databases, customer relationship management (CRM) systems, customer service logs, marketing campaign data, and external sources such as credit bureau data. This data from Customer Service may be in Excel sheets, databases, call transcripts, or customer feedback reports. The author must process the textual data and convert it to tabular data. The main domains will be Transactional, Demographic Data, Customer Interaction Data, and Product Usage Data with Customer Churn as the target feature. Some derived features that will help the analysis are minimum, maximum, and average features like credit score, age, tenure, balance number of products, and estimated salary. These derived features will help the author understand the general trend of the data.

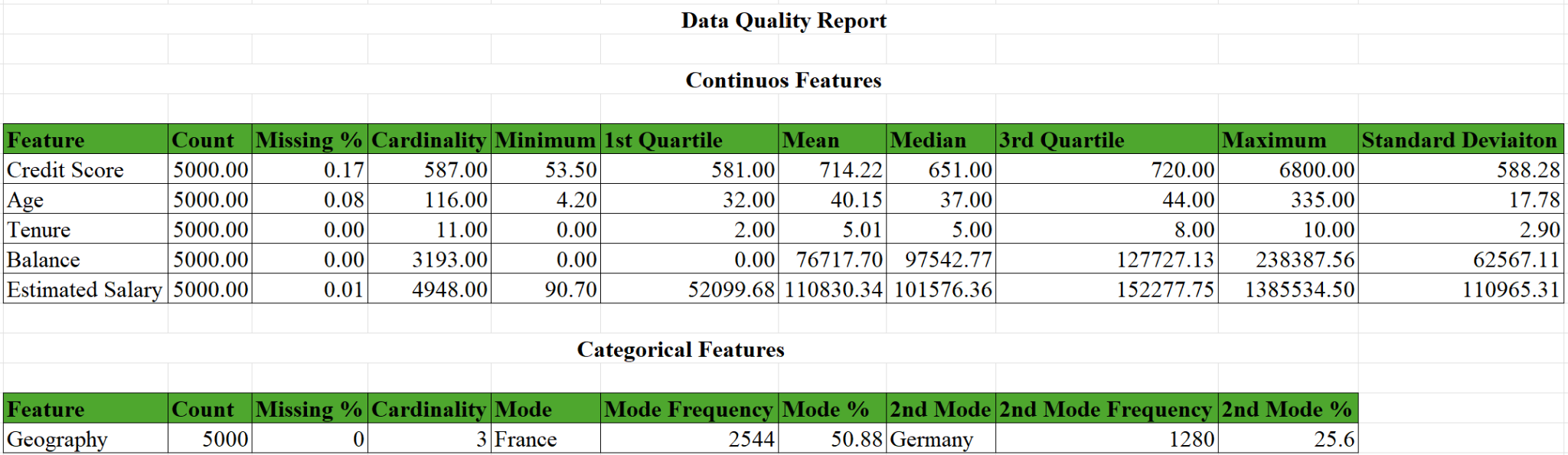
Other models include Product recommendations and Sales Prediction. The author must find these models suitable due to the lack of data and complexity while designing the ABT. The output features of these models will be non-binary. These models would also employ other statistical features like market trends, customer sentiment, product trends, and customer preferences. Such data can be complex to retrieve, making the whole process more time-consuming.

# **Practical Analysis**

For Practical Analysis, the author used only the data available to him, and he did not attempt to gather the missing data needed for better predictions. The author has used Python for arithmetic and statistical calculations. After getting the output, the author has created an analytical base table for continuous and categorical data and some graphs that explain the nature of the data.

**Table 1.**

Analytical Base Table



*Note.* From 2305505ABTcase Excel File

Table 1 describes the mathematical nature of the data. The first column is the feature name, and next is the count. This column informs readers about the number of records available for each parameter. In the authors’ data, all the features have 5000 records. The third column is the missing percentage, which informs readers about how much of the total records have no value to them. In predictive analysis, a low missing percentage is desirable as it helps improve the accuracy of predictive models. The highest missing percentage is for the credit score feature, which is 17%, followed by Age, which is 8%, estimated salary, 1%, and Tenure and Balance, 0%.

The issue of missing or null values can lower the accuracy of the prediction models. To counter the effects of this issue, we can use missing indicator and imputation techniques. In order to implement the missing indicator technique, we will have to add a new column to our dataset and input a binary flag, a numerical flag, to mark records with missing data points as 1 else 0. The prediction model will treat records with the flag as 1 with less confidence. Even though this is sufficient to improve the accuracy of the predictive model, the accuracy can be further improved by imputation, meaning we add some value in place of null records. Imputation can be done by replacing null with average value, or if we need more accuracy, we can average the neighboring observations so that the data trend can be maintained. If both are implemented, the author feels that missing indicator and imputation techniques can improve the model's accuracy.

The next column is cardinality. Cardinality is the number of unique values in a parameter present in the data. In the authors' data, the estimated salary has the highest cardinality of 4948, which means almost every past customer whose information is on record has a different salary.

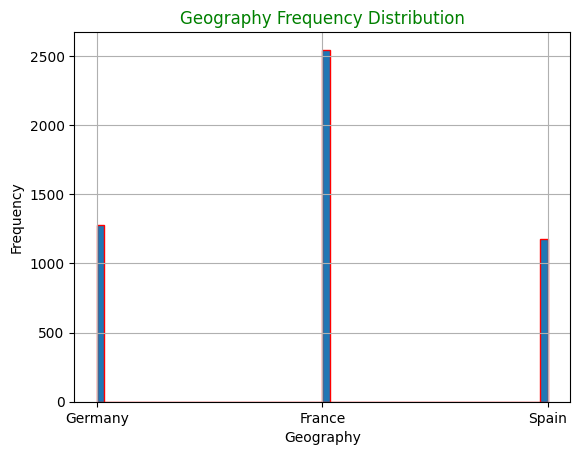
The next column shows the minimum value recorded for a particular feature; also, there is a maximum column second from the right in the table, which shows the maximum value recorded for a particular feature. The author feels that the minimum values recorded for credit score, age, and estimated salary and maximum values for credit score and age are inaccurate. The credit score value ranges from 300 to 850, and the values recorded by the author are 53 as minimum and 6800 as maximum. For age, the maximum and minimum values are improbable, as no 4-year-old child will be allowed to open an account or buy any banking product, and the maximum value is 335; no person can live for that long, and clearly is a mistake. As far as the minimum value of the estimated salary is concerned, which is 90.70, is a mistake as the minimum wage is about 16 dollars.

The 1st quartile, mean, median and 3rd quartile columns show the cutoff of 25%, the average value, the middle value, and the cutoff of 75%, respectively. These won't make much sense to ordinary readers, but the author used these values to calculate the minimum acceptable and maximum acceptable values. As their name suggests, these minimum and maximum acceptable values are the lower and upper-level cutoffs of the features, which are both statistically and logically correct and should actually be considered the logical minimum and maximum values. The minimum and maximum acceptable values for each feature are - credit score - 139 and 372.5, age - 14 and 62, tenure - -7 and 17, balance - -191590.69 and 319317.82, estimated salary - -98167.42 and 302544.85, respectively. For some features like tenure, balance, and estimated salary, the lower limits are negative, which is unacceptable as these features cannot have negative values. Such issues can be solved by filtering out the negative records or implementing the missing flag and imputation technique by treating negative values as nulls. The author prefers to remove such records

The last column shows the standard deviation value, which, in other words, is something that tells readers how much data points "deviate" or astray from the mean. A low standard deviation will show the closeness of the data points to the mean, and a high standard deviation would mean that they are more spread out. It gives readers an idea of how much variation or dispersion exists within the data set.

**Figure 1.**

Geography Frequency Distribution



*Note.* From Python Output Excel file

The next small table is the ABT for categorical features, and there is only one categorical feature: Geography. It has 5000 counts of observations, 0 missing percentages, and 3 unique observations. The next columns are related to mode, which is the most recurring observation. In the authors' data sets, the geography feature, France, is the most recurring observation, followed by Germany; Figure 1 is the graphical representation of the same. This recurrence in observation means that most of the banks' customers are from those countries. Mode frequency is the number of times a unique observation was observed, and mode percentage is the percentage of the total observations.

# **Conclusion**

In this academic adventure, we set out to decode the latent stories within real datasets to predict customer churn in the banking sector—a challenge that is felt across industries. In a quest to predict churn patterns and develop pre-emptive strategies, we set on this mission with raw data, which was much beyond just solving a problem. We refined a binary predictive model in relation to the concepts of customer churn through detailed analysis and feature engineering, overcoming the difficulty of acquisition and preprocessing of data to fine-tune the ability to predict.

In this practical analysis landscape, we faced the most prevalent issue: missing data. We bolstered the robustness of our model with innovative techniques like missing indicator flags and imputation. We documented our insights in an analytical base table, shedding light on the statistical nuances of our dataset—unveiling latent patterns and preferences that guide strategic decisions. As beacons of data-driven innovation, we stand ready to empower organizations with insights and methodologies required in order to navigate this dynamic terrain of customer retention and business resilience.

# **References**

Sghir, N., Adadi, A., & Lahmer, M. (2022, January 8). *Recent advances in Predictive Learning Analytics: A decade systematic review*. Springer Link. Retrieved May 12, 2024, from https://link.springer.com/article/10.1007/s10639-022-11536-0

Stephen, B. (2024, February 28). *(PDF) Customer Churn Prediction using Machine Learning Models*. ResearchGate. Retrieved May 12, 2024, from https://www.researchgate.net/publication/378026543\_Customer\_Churn\_Prediction\_using\_Machine\_Learning\_Models