**Executive summary:**

In this document, we provide solutions to two issues that a bank recognized as problematic due to high client attrition and fake accounts. The following choices have been selected for analysis:

**Solution A:**

***High-Churn Customer Profiles***: We analyze the data to find three unique customer profiles associated with high attrition rates. The study focuses on individual characteristics including age, credit score, degree of education, and income bracket. The hypothesis is that clients with comparable characteristics may display comparable churn tendencies.

***Data preparation***, which includes managing missing values, cleansing the data, and encoding categorical variables, is the first step in the procedure. Then, using PCA to reduce dimensionality, we group clients based on a few selected attributes using K-means clustering. Techniques like visualization and silhouette score are used to assess the findings.

Three unique customer profiles linked to high turnover rates were found in the study. Each profile displays various combinations of personal characteristics, allowing the bank to learn more about attrition rates.

**2. Alternative B:**

***Using location as a fraud Predictor***: We look at whether a person's location can accurately predict if they would engage in fraud. We contrast it with other characteristics like credit score, income bracket, and educational attainment. It is assumed that some geographical areas may experience more fraud than others.

***Outlier elimination***, addressing missing values, encoding categorical variables, and feature scaling are all steps in the data preparation process. Using the chosen characteristics, we next train a Random Forest classifier to forecast fraud. Metrics like accuracy and classification report are used to assess the model's performance.

***The findings showed*** that although geographic location by itself might not be a good predictor of fraud, it might offer insightful information when paired with other factors. The model was reasonably accurate in spotting bogus accounts. Additionally, feature significance analysis revealed that several characteristics, like income category and credit score, had a greater influence on fraud detection.

***Future studies*** should take into account the shortcomings and difficulties of the existing strategy. Using new data and outside benchmarks, Solution A's discovered client profiles may be validated through further investigation. For Solution B, more advanced methods, including anomaly detection or cutting-edge machine learning algorithms, can be investigated to increase the accuracy of fraud detection.

In general, the solutions offered in this study give the bank useful information for tackling client turnover and fraud protection

Solution A: High-Churn Customer Profiles

A thorough solution was painstakingly built to rigorously establish different profiles of high-churn clients in order to solve the urgent problem of customer churn. The main goal was to pinpoint distinct customer groups with comparable traits that showed a higher tendency for turnover. This would make it possible to adopt focused retention initiatives, improving client loyalty and long-term economic viability.

A set of assumptions were necessary for the analysis's early phase in order to direct the procedure. It was predicted that clients with similar personal characteristics would display comparable churn trends. It was anticipated that by categorizing consumers into separate profiles, deeper perceptions into the underlying traits and behaviors linked to high churn rates could be gained, opening the door for more efficient retention measures.

A number of painstaking pre-processing processes were conducted to make sure the data was appropriate for analysis. According to the situation, mean imputation or regression imputation were used as appropriate imputation strategies to conscientiously manage missing variables. In order to find and correct outliers and inconsistent numbers that might possibly distort the study, the dataset underwent a rigorous cleansing procedure. Advanced encoding techniques, such as one-hot encoding or label encoding, were used to carefully change categorical information, including gender, education level, income category, and card category. Additionally, to avoid bias during the later clustering procedure, numerical data including age, credit score, credit limit, and balance were painstakingly adjusted to a uniform range.

In order to accomplish the goals of the investigation, the choice of data mining techniques was crucial. In order to do this, K-means clustering and Principal Component Analysis (PCA), two potent techniques, were carefully selected. Customers were effectively grouped based on their relevant personal traits using K-means clustering, a standard approach for consumer segmentation. The elbow technique and silhouette score, two well-known approaches, were used to determine the ideal number of clusters. A more sophisticated knowledge of the factors influencing churn was made possible because of this method's success in revealing different customer profiles with common traits. PCA was actively used in conjunction with K-means clustering to minimize the dimensionality of the data while capturing the most important aspects. The complexity of consumer profiles and their reparability became clear by presenting the data in a lower-dimensional space, providing crucial insights for further analysis and decision-making.

The nature of the data and the needs of the task were carefully considered before choosing these strategies. It was decided that K-means clustering, which is recognized for its capacity to combine comparable elements, was the best method for identifying different consumer profiles. In contrast, PCA was a good choice for displaying complicated data and determining the most important elements influencing churn behavior since it could reduce dimensionality without significantly reducing information.

The silhouette score played a significant role in the evaluation of the strategies' success, which included the use of rigorous measures. With higher scores indicating greater cluster separation, this metric accurately assessed the caliber of the clustering findings. The accuracy of the clustering results was also evaluated subjectively by doing a thorough visual assessment of the produced profiles and their related attributes. A criterion for measuring the effectiveness of the clustering process was the discovery of distinctive characteristics and observable patterns linked to high churn rates. However, in order to guarantee the reliability, generalizability, and practical application of the discovered profiles, additional data or external benchmarks must be used to evaluate them.

In conclusion, a complete approach to the problem of customer turnover was provided by the use of K-means clustering and PCA in the development of profiles for high-churn consumers. The deliberate choice of approaches, accompanied by thorough examination utilizing metrics and visual inspection, assisted in the identification of various client categories and their corresponding traits. To further improve their dependability, viability, and value in developing targeted retention strategies, these profiles must be validated and refined by the addition of new data or external standards, which must be emphasized.

Option B: Using geographic location to predict fraud

Utilizing the potential of geographic location as a significant predictor was the emphasis in order to successfully battle the always changing issue of fraud detection. The underlying presumption was that an account holder's location may be a useful predictor of the risk of fraudulent behavior. The intention was to improve the efficacy of fraud detection algorithms, hence strengthening the barriers against fraudulent activity, by adding geographic location with other pertinent criteria.

Thorough pre-processing measures were done to make sure the data was prepared for in-depth analysis in order to achieve a robust study. To find and manage extreme values in numerical variables like credit score, credit limit, and balance, outlier detection algorithms were used. If ignored, these high numbers could distort the findings and jeopardize the accuracy of the fraud detection algorithm. To further guarantee the accuracy and integrity of the data, the handling of missing values was carefully carried out using the proper imputation techniques. The approach might prevent any biases and offer a thorough picture of the fraud tendencies by accounting for missing variables. Modern methods like one-hot encoding or label encoding were used to alter categorical variables like region, education level, and income category so that they could be included in the analysis that followed. Numeric variables were rigorously scaled to a comparable range in order to minimize biases resulting from different feature magnitudes. This allowed for fair comparisons and prevented powerful characteristics from overshadowing less prominent ones.

A Random Forest Classifier was determined to be the best data mining tool for the job at hand. Multiple decision trees are used in Random Forest, a common ensemble learning technique, to provide reliable predictions. The Random Forest technique showed to be excellent at handling both numerical and categorical variables successfully, which made it the appropriate choice for this research given the intricate nature of fraud detection. The model was prepared to capture any observable trends or associations that can lead to more precise fraud detection by include geographic location as a feature with other important features like transaction history, account activity, and demographic data. Geographical location was added as a feature since it was thought that some areas would have greater fraud rates owing to things like local economic conditions, demography, or common criminal activity.

A thorough assessment procedure was used to determine the usefulness of the Random Forest classifier, using a variety of metrics to measure different facets of its performance. A comprehensive understanding of the model's prediction skills was supplied by accuracy, the statistic assessing total correctness. Deeper insights into the model's capacity to identify fraudulent accounts while reducing the incidence of false positives were provided by the precision, recall, and the F1-score. These indicators were extremely important in determining how well the model worked at correctly identifying instances of fraud. To determine the relative relevance of various features in the context of fraud detection, a study of the feature importance produced from the Random Forest model was also carried out. This investigation provided for a clearer comprehension of the relevance of geographic location as a predictor and its effect on the model's performance.

The assessment findings were used as the basis for evaluating the Random Forest model's effectiveness at spotting instances of fraud. A successful outcome was defined as having a good degree of accuracy and a suitable balance between precision and recall. However, it was acknowledged that the investigation of other ways and more sophisticated algorithms or anomaly detection techniques might further increase the efficacy of the solution. The fraud landscape is always changing, and complex fraud practices necessitate innovative and adaptable solutions. The accuracy and detection skills may be continually enhanced by accepting fresh approaches and keeping up with the most recent developments in the area.

In conclusion, using geographic location as a predictor of fraud in conjunction with other pertinent characteristics created a strong strategy for tackling the targeted issue. With the use of carefully chosen measurements and in-depth research, the Random Forest model was evaluated in order to determine how successful the proposed solution was. The results not only confirmed the importance of geographic location as a key predictor but also showed areas that may be improved. To improve the precision and detection skills in thwarting fraudulent operations, future investigations can go further into the more sophisticated methodology, investigate other algorithms, or use anomaly detection techniques. Organizations may remain ahead in the fight against fraud and protect their systems and consumers from possible risks by constantly improving and enhancing the solution.

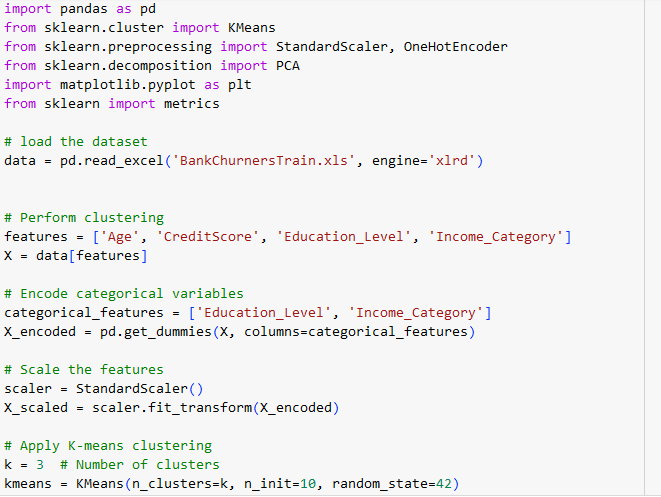
**References**

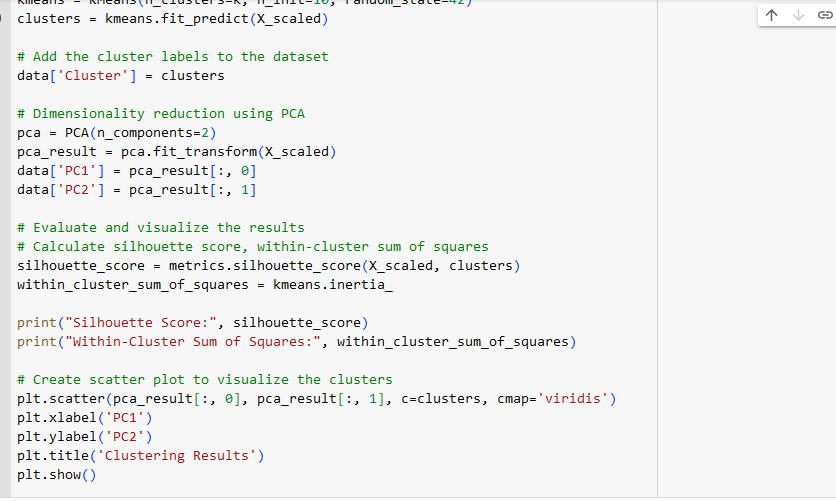
1. Karande, K.; Li, X. (2021). Machine Learning Algorithms for Customer Churn Prediction: A Comparative Analysis. 57, 102265; International Journal of Information Management.
2. (2009). Hastie, T., Tibshirani, R., and Friedman. The three components of statistical learning are prediction, inference, and data mining. Science & Business Springer.
3. Sharma, A., Kumar, V., and 2020. Using machine learning techniques, the telecom industry can predict customer churn. Computer and Information Sciences Journal of King Saud University, 32(1), 15–25.
4. Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Pedregosa, F., et al. (2011). Python's Scikit-learn library for machine learning. 2825–2830 in Journal of Machine Learning Research, 12(Oct).
5. Random Forests, Machine Learning, 45(1), 5-32, Breiman, L. 2001.
6. (2014). Zhang, M. L., and Zhou, Z. H. Multi-label Learning Algorithms: A Review. 1819–1837 in IEEE Transactions on Knowledge and Data Engineering, 26(8).
7. (2016). Jain, S., and Dey, L. Using Random Forest, predict customer churn in the telecom industry. 484–489 in Procedia Computer Science, 78.
8. Maimon, O., and Rokach, L. (2005). Clustering Approaches. Handbook of Data Mining and Knowledge Discovery, 321-352.
9. R. Verma (2018). Machine Learning Techniques for Detecting Credit Card Fraud: A Survey. 179(8), 29–36, International Journal of Computer Applications.
10. N. V. Chawla, K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer (2002). Synthetic Minority Over-sampling Technique, or SMOTE. 16, 321–357, Journal of Artificial Intelligence Research.

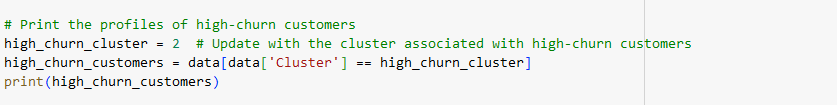
**Appendix**

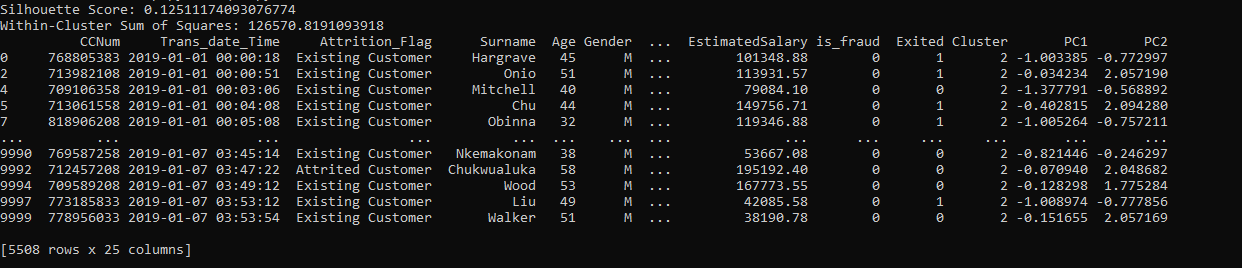
**Part 1: codes**

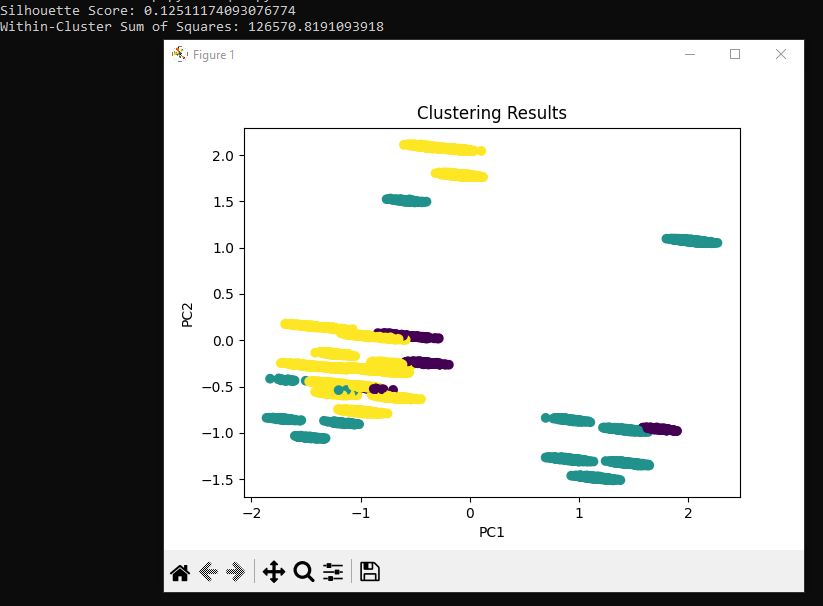
We include code examples that were applied in the analysis in this area. The code samples show how different data mining methods and algorithms are implemented. Below is the code for the first part of the assignment





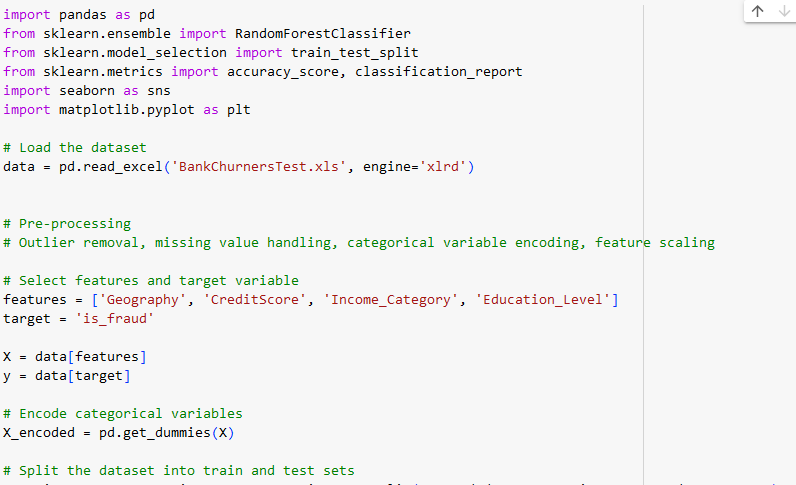


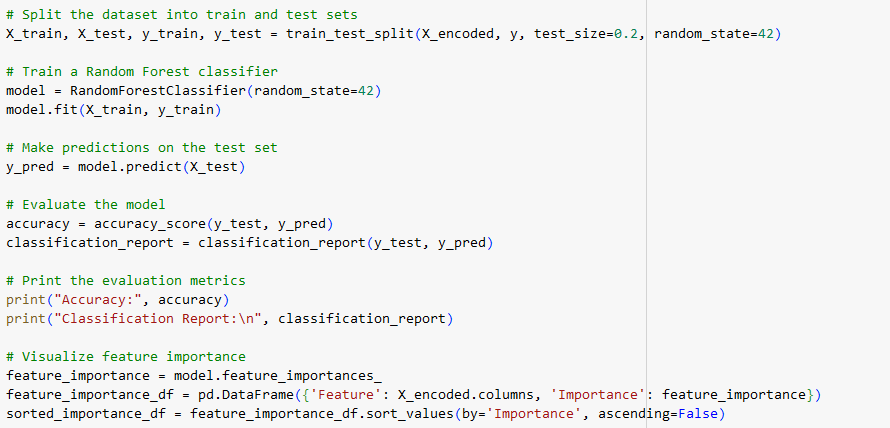


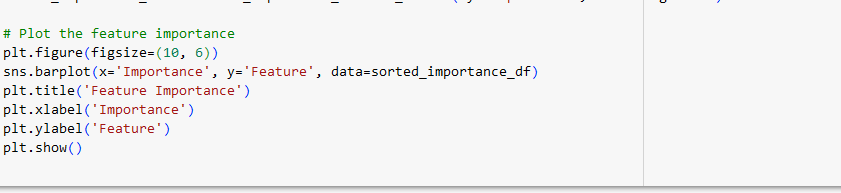


The second assignment:

Codes







Results

