**Project 2 - Revisited**

For project 2, you were given a masked data set and tasked with building a predictive model that produced the highest possible recall (while trying to balance recall and the f1 score). The focus of the exercise was entirely on prediction and not on interpretation.

Before taking a look at the information on the following pages, please answer the following question:

**Question 1:**

1. What did you learn from this exercise about the technical process of building a machine learning model on a large unbalanced data set?

**Key takeaways from this exercise on building a machine learning model with a large, unbalanced dataset include the critical importance of addressing data imbalance early in the process. Initially, our model was developed without balancing the dataset, but after implementing SMOTE in our second iteration, we observed significant improvements, particularly in our F1 score and recall. Additionally, we learned that sampling a subset of the data can be beneficial for gaining familiarity with the modeling process and efficiently testing multiple techniques to identify the most optimal approach in a shorter timeframe.**

1. What additional information do you wish you would have had before you started?

**As a team, we would have benefitted from understanding that we were dealing with data that directly correlated to profit. We also learned in class this week that precision and recall do not always correlate with the best financial outcomes, this also would have been a helpful tip if we knew we were working directly with profit predictions. It would have been helpful to know the business’s expectations for model performance, like an acceptable F1 score range, and how much data manipulation (like SMOTE) was allowed. Knowing the meanings of the columns would have also guided more effective feature engineering and possibly simplified the model.**

1. What other resources would you have liked to have access to while building the model?

**Having access to the company’s previous models would have provided useful benchmarks, and a more powerful computer would have significantly reduced the time spent running optimization processes on the large dataset.**

1. (Feedback for the professor – please be honest) Do you believe this was a valuable exercise? Why or why not? If this exercise was used in the future, are there any modifications that you would suggest?

**We found this exercise to be incredibly valuable. While working in an office setting, we typically have a clear understanding of the research topic and the models we are developing. However, we may not always have access to the most relevant, high-quality, or up-to-date information. This exercise provided a great opportunity to practice adapting to such challenges and refining our approach accordingly. We enjoyed the hands-on experience of testing different models and optimizations and found it both educational and fun. It also gave us a great project to showcase on the resume and GitHub.**

The dataset that you analyzed in project 2 was designed to simulate a realistic bank account fraud scenario. Bank account fraud refers to any deceptive or illegal activity intended to gain unauthorized access to a person's or organization's bank account. This can include stealing money, making unauthorized transactions, or using stolen credentials to manipulate account information. For each observation in the data set, there is a variable indicating whether or not the account activity was fraudulent, and there are 31 features that are commonly related to bank account fraud. Information about each variable is provided in the table below:

**Data Dictionary:**

| **Masked Variable Name** | **Original Variable Name** | **Variable Description** |
| --- | --- | --- |
| Target\_Y | fraud\_bool | Fraud label (1 if fraud, 0 if legit) |
| X1 | Income | Annual income of the applicant in quantiles (ranges between 0 and 1) |
| X2 | Name\_email\_similarity | Metric of similarity between email and applicant’s name. Higher values represent higher similarity. |
| X3 | Prev\_address\_months\_count | Number of months in previous registered address of the applicant |
| X4 | Current\_address\_months\_count | Months in currently registered address of the applicant. (-1 is missing) |
| X5 | Customer\_age | Applicant’s age in bins per decade (e.g., 20-29 is represented as 20) |
| X6 | Days\_since\_request | Number of days passed since application was done |
| X7 | Intended\_balcon\_amount | Initial transferred amount for application |
| X8 | Payment\_type | Credit payment plan type (5 possible values) |
| X9 | Zip\_count\_4w | Number of applications within same zip code in last 4 weeks |
| X10 | Velocity\_6h | Velocity of total applications made in last 6 hours |
| X11 | Velocity\_24h | Velocity of total applications made in last 24 hours |
| X12 | Velocity\_4w | Velocity of total applications made in last 4 weeks |
| X13 | Bank\_branch\_count\_8w | Number of total applications in the selected bank branch in the last 8 weeks |
| X14 | Date\_of\_birth\_distinct\_emails\_4w | Number of emails for applicants with the same date of birth in the last 4 weeks |
| X15 | Employment\_status | Employment status of the applicant (7 possible values) |
| X16 | Credit\_risk\_score | Internal score of application risk |
| X17 | Email\_is\_free | Domain of application e-mail (free or paid) |
| X18 | Housing\_status | Current residential status (7 possible values) |
| X19 | Phone\_home\_valid | Validity of provided home phone number |
| X20 | Phone\_mobile\_valid | Validity of provided mobile phone number |
| X21 | Bank\_months\_count | How old is previous account (if held) in months. -1 is missing |
| X22 | Has\_other\_cards | If applicant has other cards from the same banking company |
| X23 | Proposed\_credit\_limit | Applicant’s proposed credit limit |
| X24 | Foreign\_request | If origin country of request is different from bank’s country |
| X25 | Source | Online source of application |
| X26 | Session\_length\_in\_minutes | Length of user session in banking website in minutes |
| X27 | Device\_os | Operating system of device that made request |
| X28 | Keep\_alive\_session | User option on session logout |
| X29 | Device\_distinct\_emails\_8w | Number of distinct emails in banking website from the used device in last 8 weeks |
| X30 | Device\_fraud\_count | Number of fraudulent applications with used device |
| X31 | Month | Month when the application was made |

**Question 2:**

Using your best and final model, apply the techniques that you have learned around explainable machine learning to identify the top five features associated with bank account fraud. List the features below with a brief explanation of how each features impacts bank account fraud.

**Utilizing our best and final model, along with the techniques we have learned from explainable machine learning, we have identified the five key features most indicative of bank account fraud:**

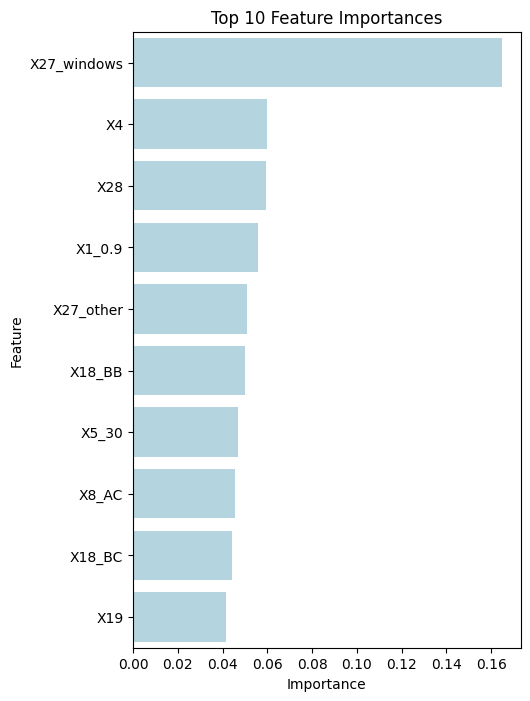
**Device Windows: The operating system used for the request, Windows has the highest correlation of fraud (take into account how many individuals are Windows owners)**

**Current Address (Months of Tenancy Count): The duration an applicant has lived at their current address, where shorter periods may indicate higher fraud risk**

**Session Alive (Keep Session Alive): Whether a user opts to keep their session active or alive, this can be linked to odd and suspicious behaviors**

**Income (Income - Top Quantile): Applicants with incomes in the highest quantile which could be exploited further in fraudulent activity**

**Device Other: Requests from less common operating systems may signal potential fraud attempts**

****

Suppose that you were also given the following information regarding the costs associated with bank account fraud and the potential benefit of proactively identifying fraudulent activity. Your endgame has now changed, and the bank wants you to develop a model that will maximize profit (or minimize cost) associated with bank account fraud.

Note that total profit could be computed as:

Profit = (TP\* gain per TP) – (FN \* cost per FN) – (FP \* cost per FP)

| **Scenario** | **Description** | **Estimated Cost ($)** | **Estimated Benefit ($)** |
| --- | --- | --- | --- |
| **True Positive (TP)** | Correctly identifying fraudulent activity in a bank account and preventing it. | **Cost**: Minimal (monitoring, fraud investigation, customer notification). | **Benefit**: $1,000–$100,000+ (prevention of unauthorized withdrawals, identity theft, legal fees, and reputational damage). |
| **False Negative (FN)** | Missing a fraudulent transaction or account takeover, allowing fraud to occur. | **Cost**: $1,000–$500,000+ (stolen funds, compliance penalties, customer compensation, legal fees, reputational harm). | **Benefit**: None (except temporarily avoiding fraud investigation costs). |
| **False Positive (FP)** | Incorrectly flagging a legitimate transaction or account activity as fraud, causing inconvenience to the customer. | **Cost**: $100–$5,000 (customer frustration, operational costs, potential account closure or loss of business). | **Benefit**: Minimal (a potential deterrent effect on actual fraudsters). |
| **True Negative (TN)** | Correctly identifying legitimate account activity as safe. | **Cost**: $0 | **Benefit**: Smooth banking experience, customer satisfaction, and operational efficiency. |

**Question 3:** What is the estimated profit for the best and final model that you created? Be sure to state the values you used for cost and gain. Note that you can use the confusion matrix for the training data from your best and final model to help answer this question.

**Confusion Matrix:**

|  | **Predicted** | | |
| --- | --- | --- | --- |
| **Actual** |  | **0** | **1** |
| **0** | **TP: 296,680** | **FP: 3,085** |
| **1** | **FN: 281** | **TN: 293,337** |

**Gain per TP: $50,000**

**Cost per FN: $250,000**

**Cost per FP: $2,500**

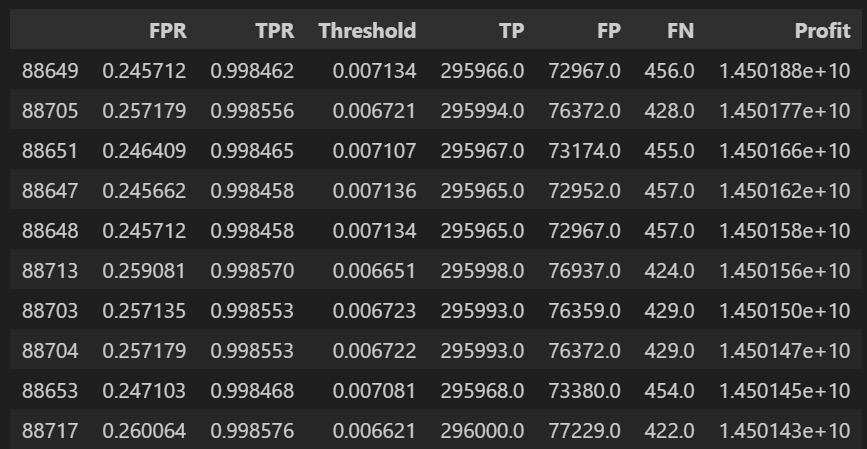
**Profit = (293,337 \* 50,000) – (281 \* 250,000) – (3,085 \* 2,500)**

**Profit = $14,588,887,500**

**The expected profit for the best and final model is $14.6 billion, based on conservative mid-range estimates for costs and gains to provide a balanced and realistic assessment.**

**Question 4:** Continuing with the same scenario from question 3, what threshold cutoff would you choose to maximize profit (based on your best and final model)? What is the value of profit for this threshold cutoff?

**To maximize profit, we selected a threshold of 0.0071, resulting in an expected profit of $14,501,880,000. This threshold was chosen as it effectively balances capturing the maximum number of fraudulent cases while minimizing the financial impact of false positives and false negatives. Using conservative cost and gain estimates, this approach prioritizes business value over traditional accuracy metrics, aligning the model’s output directly with profit maximization goals.**



**Question 5:** What might you have changed about how you approached the modeling process now that you have seen the full information about the variables and you have been told that maximizing profit is the end goal?

**Given the full information about the variables and the emphasis on maximizing profit, we would have taken a more strategic, profit-driven approach rather than focusing solely on classification accuracy. Instead of treating all errors equally, we would have implemented a cost-benefit matrix to properly weigh the impact of false positives and false negatives, ensuring that our model prioritized recall where it mattered most.**

**Additionally, we would have explored customer scoring functions to refine predictions and tailor decision-making based on potential business impact. Considering the nature of the problem, regression might have been a better approach than classification, allowing for more granular predictions. Furthermore, we would have leveraged Cost-Sensitive Learning to minimize the more costly errors, ensuring the model aligned with real-world business objectives. Finally, A/B testing would have been a key validation step, enabling us to compare model performance in a controlled environment and directly measure its impact on profitability.**

**Additionally, we would have placed a greater emphasis on threshold optimization earlier in the modeling process. Rather than defaulting to a 0.5 cutoff, we would have actively explored various thresholds from the beginning, aligning the model's sensitivity with profit maximization goals. By continuously evaluating the trade-offs between true positives, false positives, and false negatives at different thresholds, we could have identified the most profitable decision boundary sooner, ultimately improving the model's business impact.**