Fraud Detection Case Study

By Yuhe Tian

Table of Content

1. Context

1. Context

- Our Business:
  - We are a FinTech company that provides payment services to small businesses, onboarding them and opening accounts to receive funds from their operations.
- Why Fraud Detection Is Important?
  - Fraudsters steal business owners' IDs to receive payments or take over existing accounts. Our company is liable for these losses, making it crucial to detect and reduce fraud.
- Problems to Solve:
  1. Develop key metrics to monitor the risk levels of payment applications, identify anomalies, and determine their causes.
  2. Evaluate the potential benefits of partnering with the third-party vendor FraudKiller, which can provide additional fraud detection data at a certain cost.

2. Anomaly Detection

Develop key metrics to monitor the risk levels of payment applications, identify anomalies, and determine their causes.

2.1 Data Collection

- The dataset consists of payment application records used to detect and analyze fraud. It contains 233,839 applications, each represented by 11 features that capture various aspects of the applications.

Column	Description
application_id	unique ID for each individual payment application
product	Product type
industry	which industry the merchant's business belongs to
city	city where the business is located
state	state where the business is located
application_date	date when the payment application is submitted
final_decision	final decision of our onboarding process for the payment application
is_fraud	flag to indicate whether the application is fraud application
credit_score	credit score provided by 3rd party vendor
fraud_score	fraud score generated from our in-house model
first_transaction_date	date of the first transaction date from the merchant

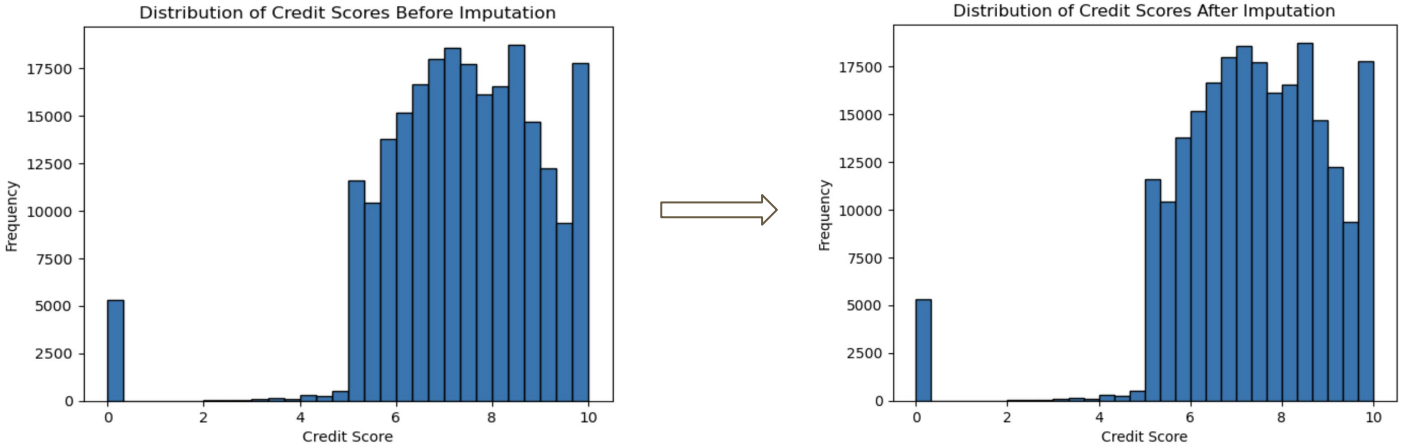
2.2 Data Cleaning - Overview

- The ‘application_id’ column was checked for uniqueness, confirming no duplicate rows. The ‘product’ and ‘industry’ columns were then examined, revealing that both contained valid and expected values. The ‘city’ and ‘state’ columns were also inspected, with unique values checked for consistency.
- An initial check revealed the number of missing values in each column: ‘credit_score’ (14,071), ‘fraud_score’ (5,447), and ‘first_transaction_date’ (129,095).

```
application_id              0
product                     0
industry                    0
city                        0
state                       0
application_date            0
final_decision              0
is_fraud                    0
credit_score            14071
fraud_score              5447
first_transaction_date  129095
dtype: int64
```

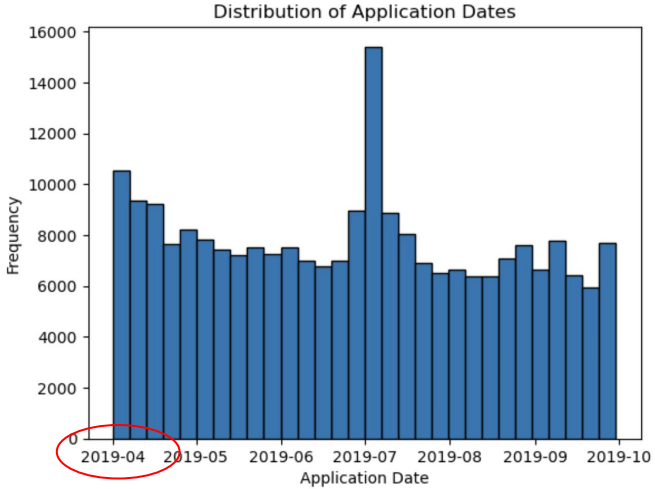
2.2 Data Cleaning - Missing Values

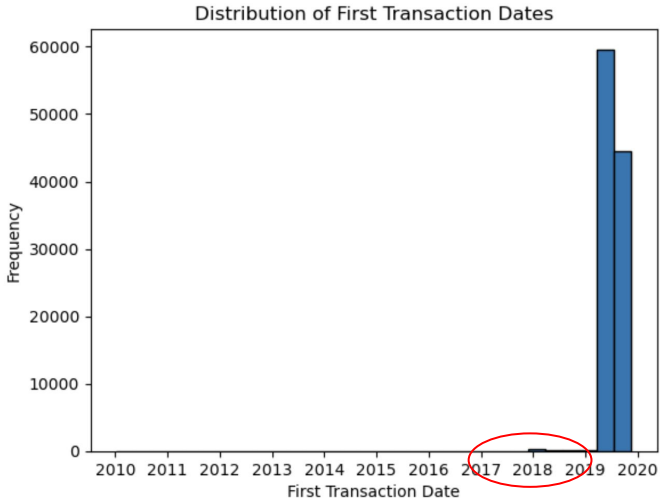
- KNN imputation is applied to address missing values in ‘credit_score’ and ‘fraud_score’ column
  - The overall distribution of the original variables are preserved.
- Missing values in ‘first_transaction_date’ shall not be imputed
  - They make sense if the user hasn’t made their first transaction yet.



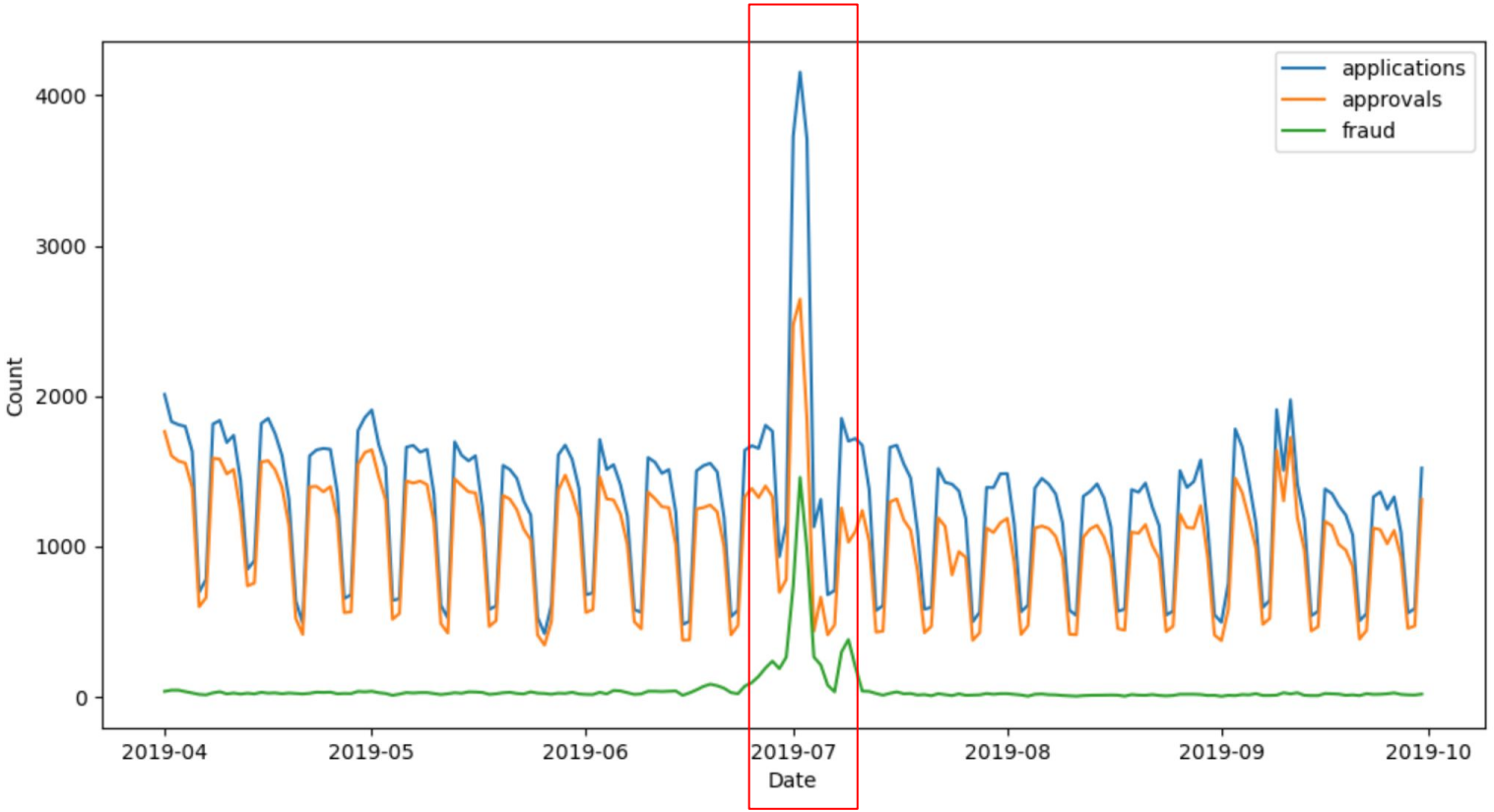
2.2 Data Cleaning - Anomalies

- Anomalies in the ‘first_transaction_date’ column were identified.
  - Some dates were recorded as earlier than the ‘application_date’, indicating a recording error.
  - These anomalies were replaced with their corresponding application_date.

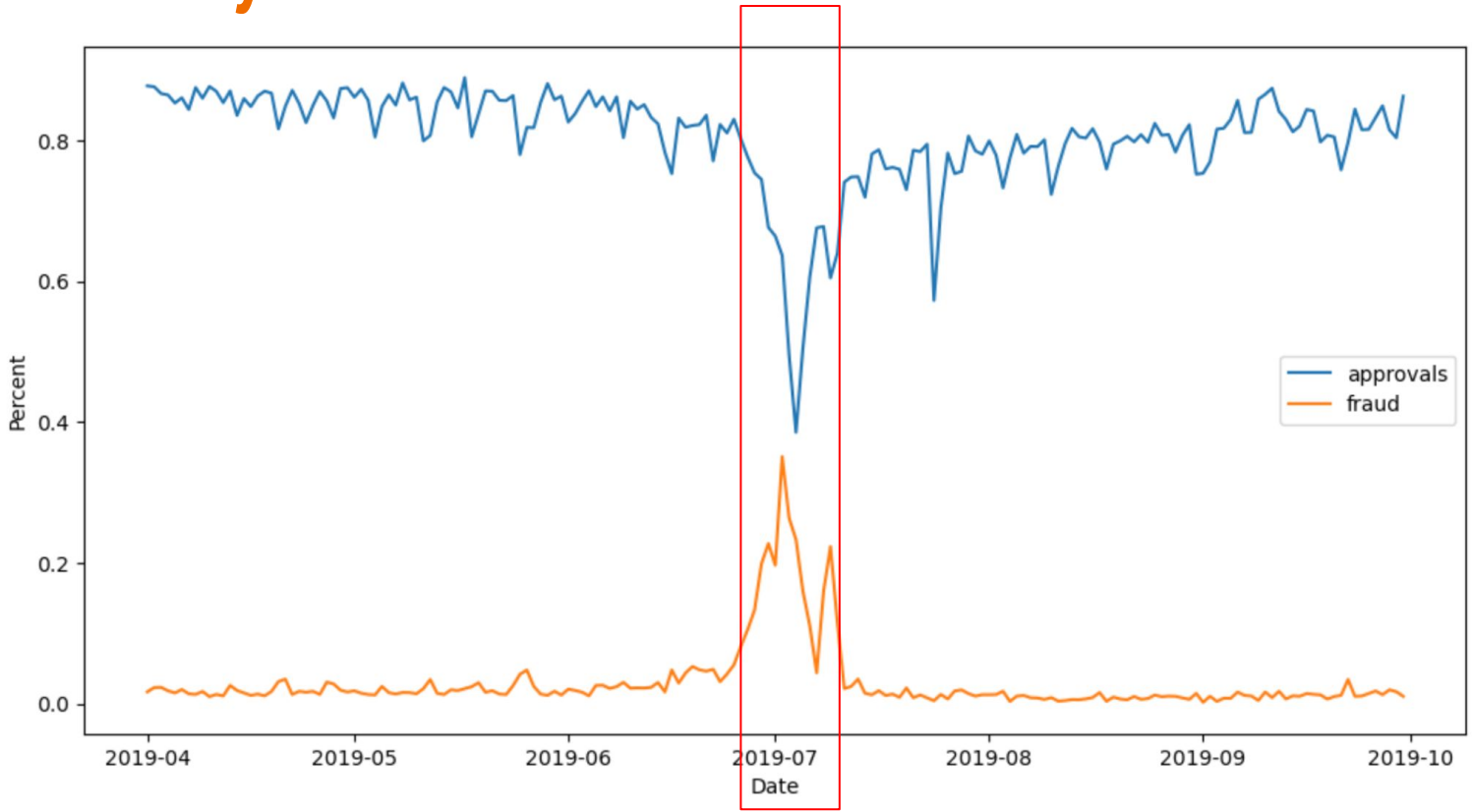




2.3 Anomaly Detection

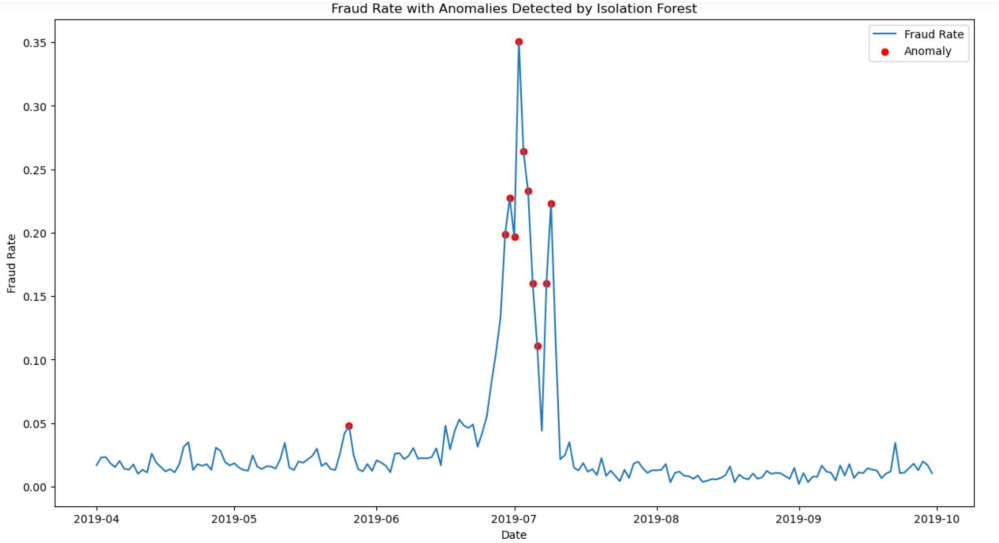


2.3 Anomaly Detection

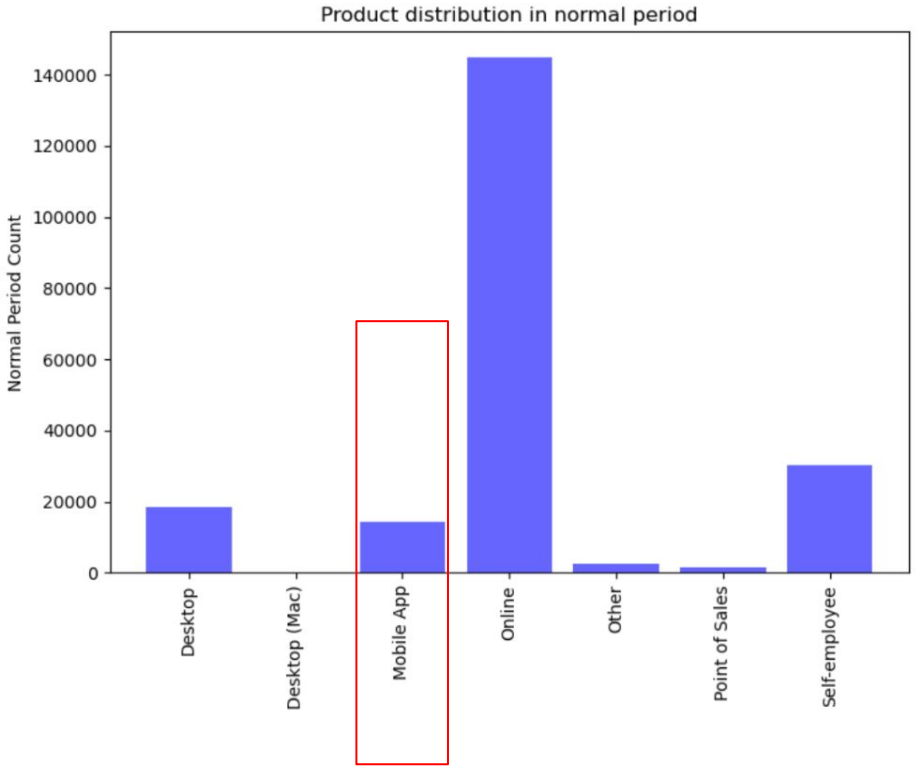


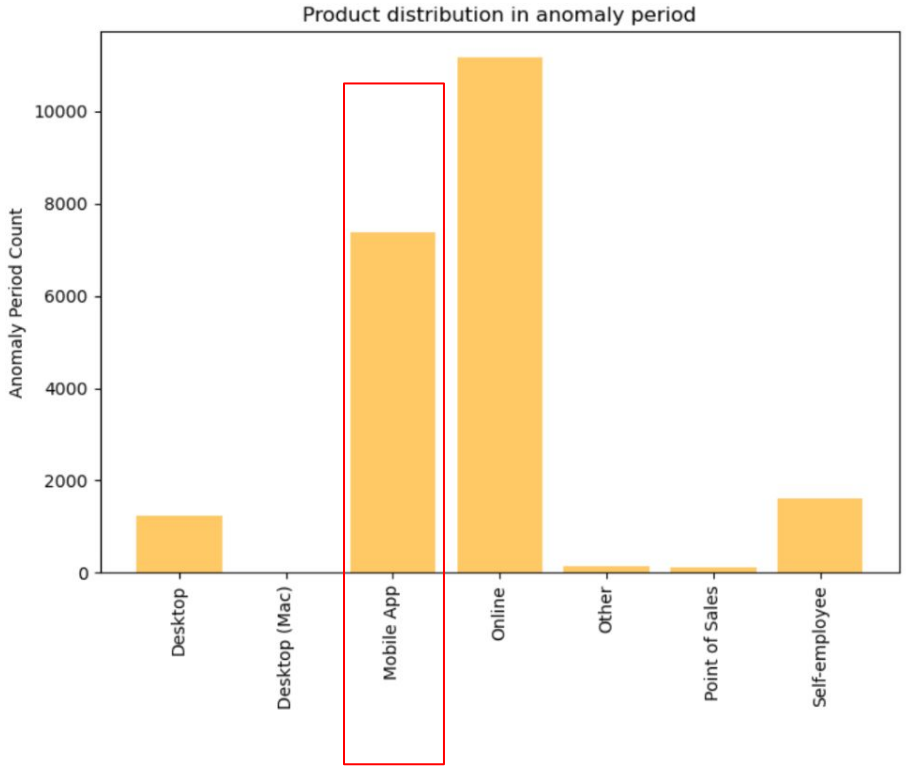
2.3 Anomaly Detection - Isolation Forest

- Why Isolation Forest?
  - Specifically designed to identify anomalies
  - Can detect complex patterns of fraud that might not be apparent when looking at individual features alone
  - Robust to high-dimensional data and does not assume any specific data distribution
- Results suggest that anomalies happened around 6/27 - 7/9

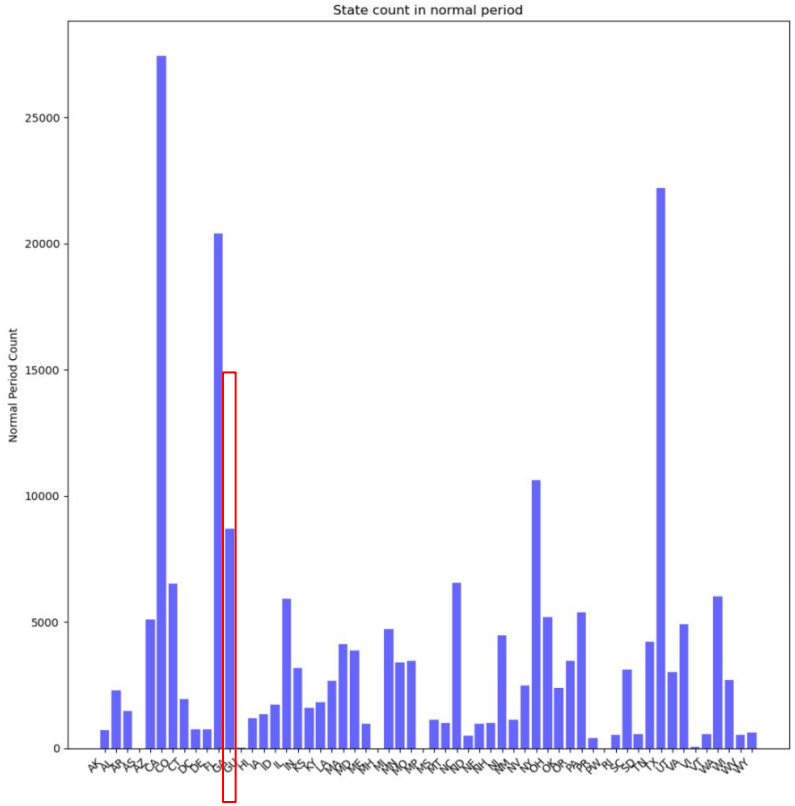


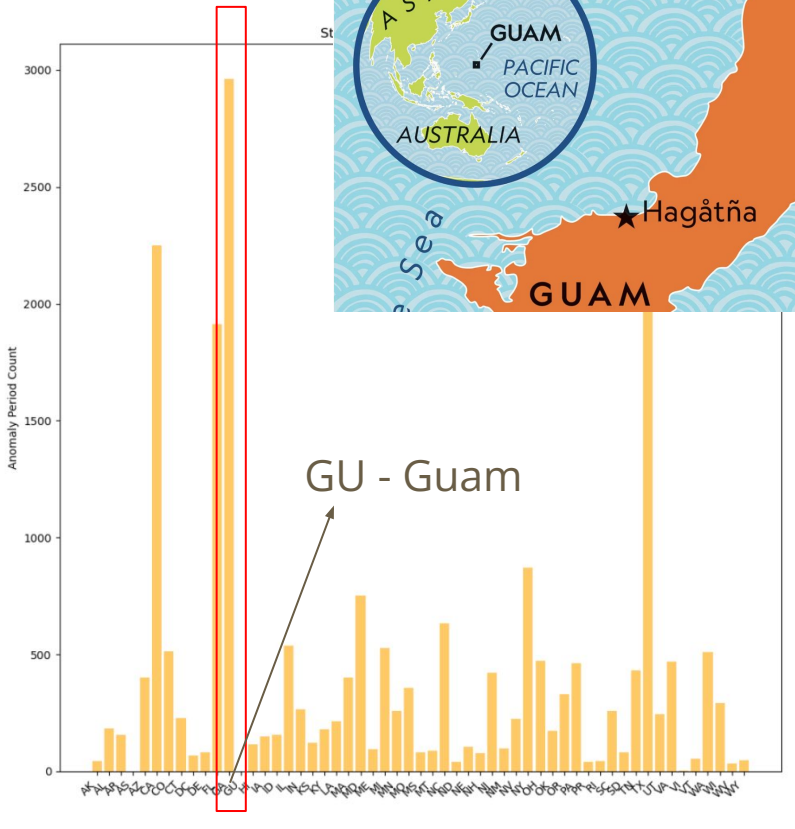
2.4 Root Cause Analysis





2.4 Root Cause Analysis





2.4 Root Cause Analysis

- The spike in fraud rates could be due to a targeted fraud campaign. Fraudsters might have exploited vulnerabilities in the system and specifically targeted mobile apps businesses and businesses located in Guam. This targeted approach indicates a strategic effort to bypass security measures and capitalize on specific weaknesses in the system.
    - Less Sophisticated Security Infrastructure
    - Smaller businesses, less information, less rigorous identity check.
    - ...

3. Should we work w/ a third party vendor FraudKiller?

Evaluate the potential benefits of partnering with the third-party vendor FraudKiller, which can provide additional fraud detection data at a certain cost.

3.1 Data Collection

- The dataset comprises 2,775 rows and 18 columns, which are divided into two categories: existing data and data provided by the third-party vendor, FraudKiller.

Existing Data	
ID	Unique Customer Identifier
is_fraud	This is a yes/no flag indicating whether an acocunt was verified as a fraud takeover (1=fraud, 0=ok)
opendate	Date the account was opened
AreaCode	Phone area code of the primary account holder
EAScore	Score ranking the riskiness of customer's email address
IdentityRank	Ranking of the fraud likelihood of the customer's identity
DeviceBrowserType	What internet browser was the customer using to apply
IpAddressLocCity	What city was the IP address linked to
IpAddressLocCountry	What country was the IP address linked to
Data Provided by the Third Party Vendor FraudKiller	
IsValid	Does FraudKiller think the device is valid
IsConnected	Does FraudKiller think the device is connected
PersonalDevice	Is the device a personal device
Reputation Level	FraudKiller reputation ranking
ReceivingMail	Does the device receive mail
Type	FraudKiller has assigned a device type to each record
Volume Score	FraudKiller assigns a volume score to each record
Result Number	How many results were returned for this record
EmailDays	Number of days since the customers email address was created

3.2 Feature Engineering

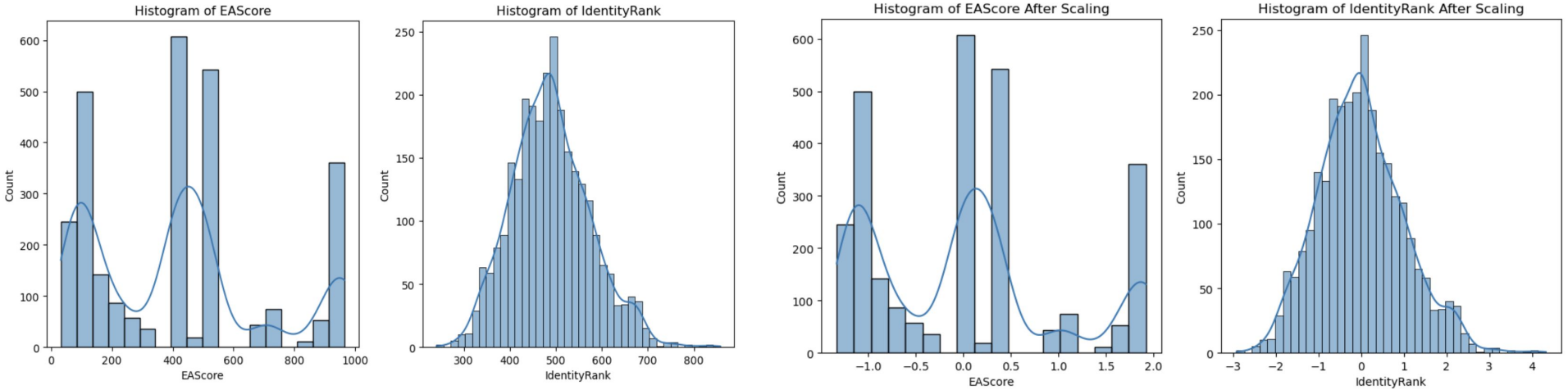
City
New York
San Fransisco
Los Angeles
Dallas

City	City
New York	2
San Fransisco	3
Los Angeles	1
Dallas	0

- Feature engineering is the process to map raw variables into features that could be used for ML modeling
- Feature Creation (Creating new features from existing features)
  - ‘IPAddressLocCity_is_null’: Indicates if the IPAddressLocCity is null.
  - Extracted ‘hour’, ‘day_of_week’, and ‘week_of_year’ from the ‘opendate’ variable.
- Feature Encoding (Converting categorical variables to values that models could process)
  - One-hot encoded categorical features such as AreaCode and DeviceBrowserType.

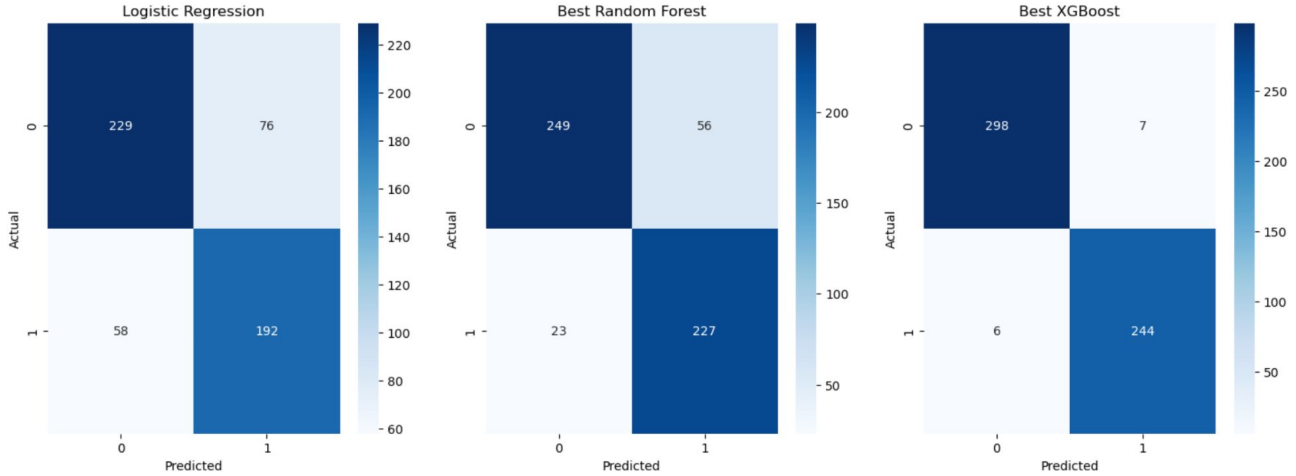
3.2 Feature Engineering

- Feature Scaling (Uniform or similar scaling with different features)
  - Standardized numerical features ('EAScore', 'IdentityRank') to have a mean of 0 and a standard deviation of 1.



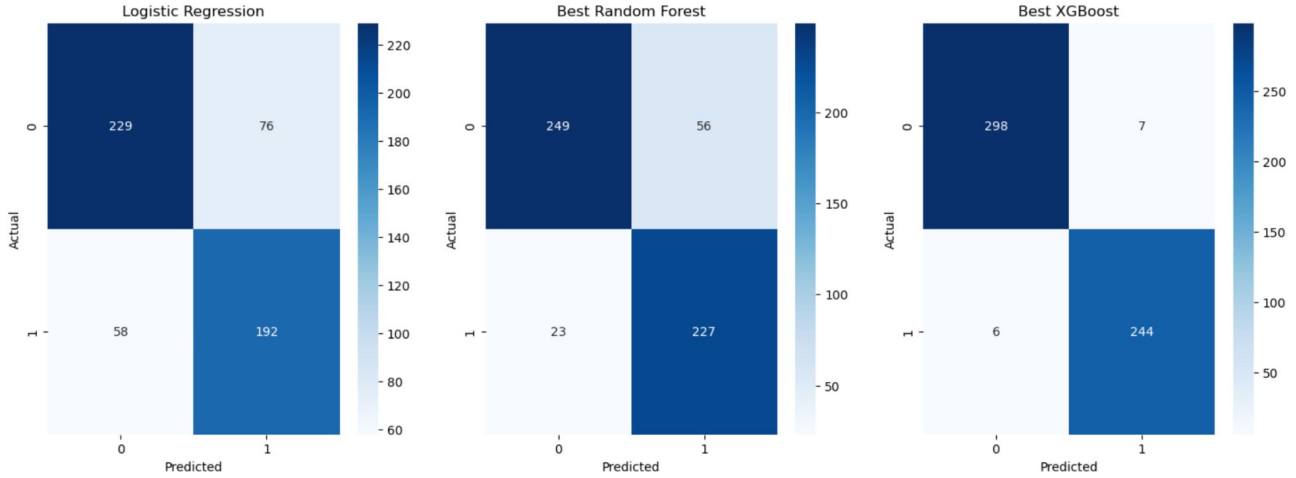
3.3 Model Selection - Without FraudKiller

- Splitted the dataset with only the existing variables into training (80%) and testing (20%) set
- Built a baseline model using Logistic Regression
- Then trained a Random Forest model with hyperparameter tuning
- The third model is an XGBoost model with hyperparameter tuning as well



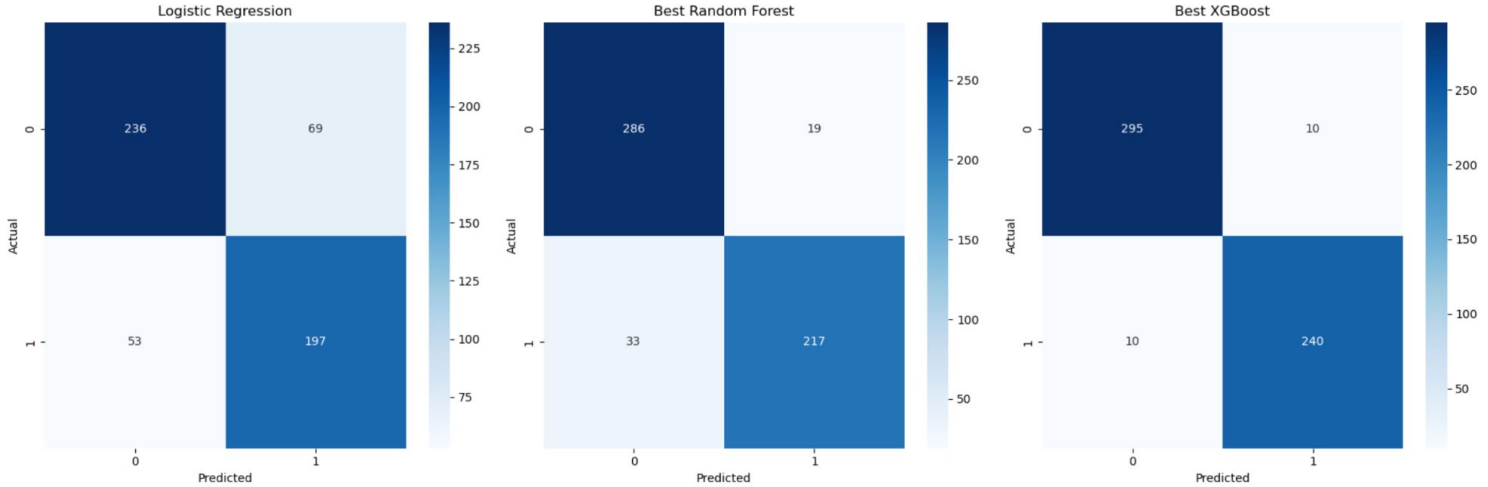
3.3 Model Selection - Without FraudKiller

- XGBoost has the highest number of True Positives and True Negatives, and the lowest number of False Positives and False Negatives, indicating it is performing the best in terms of correctly identifying both classes.
- The confusion matrix for XGBoost shows the least misclassifications, further supporting that XGBoost is the best model among the three.



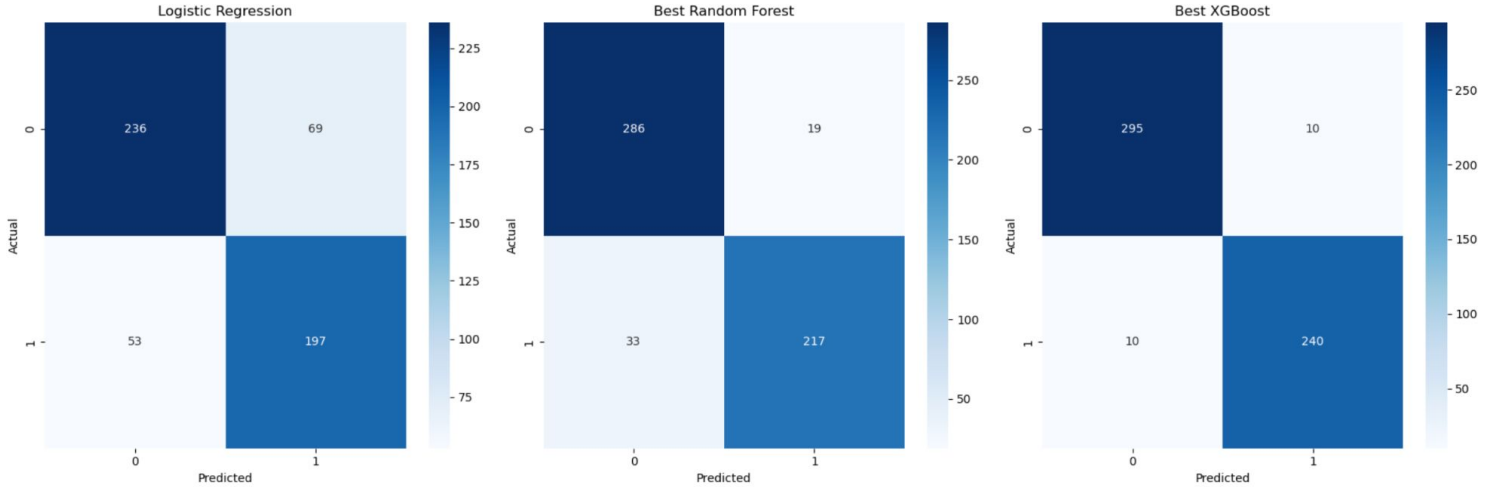
3.3 Model Selection - With FraudKiller

- Splitted the dataset with existing variables and information that could be purchased from FraudKiller into training (80%) and testing (20%) set
- Built a baseline model using Logistic Regression
- Then trained a Random Forest model with hyperparameter tuning
- The third model is an XGBoost model with hyperparameter tuning as well



3.3 Model Selection - With FraudKiller

- In this case, XGBoost still has the highest number of True Positives and True Negatives, and the lowest number of False Positives and False Negatives, indicating it is performing the best in terms of correctly identifying both classes.
- The confusion matrix for XGBoost shows the least misclassifications, further supporting that XGBoost is the best model among the three.



3.4 Profit Optimization

- To see maximum profit yielded by each model, the objective function is defined as follows:
  - Profit=Revenue+Fraud Loss+Manual Review Cost+Vendor Cost
  - Where:
    - Revenue=$40\times12\times\text{count}(\text{approval}\&(\text{not fraud}))+40\times0.3\times12\times\text{count}(\text{manual review}\&(\text{not fraud}))$
    - Fraud Loss=$-500\times\text{count}(\text{approval}\&\text{fraud})-500\times0.3\times\text{count}(\text{manual review}\&\text{fraud})$
    - Manual Review Cost=$-50\times\text{count}(\text{manual review})$
    - Vendor Cost=$-0.5\times\text{num\_data}$
  - With Constraints on threshold [x1, x2] (x1: threshold between decline & MR, x2: threshold between MR & approval):
    - $0.02\leq x1\leq0.98$ & $0.02\leq x2\leq0.98$
    - $x2-x1\geq0.05$

3.4 Profit Optimization

- If we do not purchase data provided by FraudKiller:
  - The optimal thresholds are 0.48019951 and 0.65016786.
  - The maximized profit is 139244.5.
- If we purchase data provided by Fraudkiller:
  - The optimal thresholds are 0.29998906 and 0.39998711.
  - The maximized profit is 136936.5.
- Turns out that purchasing data from the third party FraudKiller would not increase the maximum profit that we can reach

4. Conclusion

4. Conclusion

- It seems surprising that the model without FraudKiller data yields a slightly higher profit compared to the model with FraudKiller data. That may due to several reasons:
  - Quality Over Quantity: The quality of the additional data might not be high. If the FraudKiller data contains a significant amount of noise or inaccuracies, it can negatively impact the model's performance.
  - Model Complexity: Incorporating more data can increase the complexity of the model, making it harder to optimize and interpret.
  - Initial Model Strength: The initial model without the FraudKiller data is already highly optimized. In such cases, the room for improvement is limited.

4. Conclusion

- Here are some main takeaways from this project:
  - During the anomaly detection phase, we identified a spike in fraud rates. This campaign likely exploited vulnerabilities within the system, focusing on businesses in Guam or mobile app businesses. These insights highlight the importance of continuous monitoring and updating of our security measures to adapt to evolving fraud tactics.
  - The findings imply that it might be more cost-effective to focus on optimizing and enhancing the existing model using the current dataset. This approach ensures that resources are allocated efficiently without incurring unnecessary costs.

4. Conclusion

- Potential Issues and Room for Improvement:
  - Data Collection: The current dataset might not be large enough to build a model that generalizes well to future records. Expanding the dataset with more samples will help create a more robust and accurate model.
  - Features Collection: While the current model performs well, there is always room for improvement. Collecting additional high-quality, relevant features and keep updating feature selection could enhance the model's predictive power.
  - Model Precision: To improve model precision, employing advanced techniques such as ensemble learning, deep learning, and unsupervised learning algorithms can further refine the model’s accuracy and robustness.

Thanks!