DSCI445 Term Project - Bank Account Fraud Detection

Nick Brady, Jakob Wickham, Noah Sturgeon

December 2, 2024

0.1 Introduction

0.1.1 Motivation

Fraud detection in banking is a difficult problem to tackle for several reasons. Firstly, there is far more data for safe transactions than fraudulent ones. Fitting models without some care for the imbalance in the data can result in a model that cannot detect any fraud. Secondly, false negatives are far more costly than false positives. A false negative will incur losses for the bank (possibly far more than what could have been gained from a single customer), while a false positive only loses a single customer's business.

Our goal in this project is to work with various machine learning methods to get the highest recall (how many fraudulent transactions are successfully identified) possible, while locking our false positive rate (how many valid transactions are misidentified as fraud) at 5% or lower. The reason for using recall, as opposed to accuracy, is that the afore-mentioned data imbalance means that aiming for accuracy will create a model that is useless at detecting fraud.

0.2 Dataset overview

The Bank Account Fraud NeurIPS 2022 datasets (called BAF for short) are a suite of synthetic datasets meant to evaluate machine learning methods. The datasets contain 1000000 bank account transaction instances, and 31 features per transaction. Each transaction in BAF was also generated with a fraud label, a boolean value called fraud_bool, which is the response on which machine learning methods can be trained on.

The 31 features used in the dataset are:

Categorical

- **`payment_type`** - Anonymized credit payment plan type (5 types: AA, AB, AC, AD, AE)
- **`employment_status`** - Anonymized employment status of the applicant (7 types: CA - CG)
- **`housing_status`** - Anonymized residential status of the applicant (7 types: BA - BG)
- **`source`** - Online source of the application; INTERNET or TELEAPP
- **`device_os`** - Operating system of the device (e.g., Windows, macOS, Linux, X11, other)

Numeric

- **`income`** - Annual income of the applicant in quantiles; x [0.1, 0.9]
- **`name_email_similarity`** - Similarity between email and applicant's name; ranges from [0, 1]
- **`prev_address_months_count`** - Months at previous address; ranges from [-1, 380] (-1 = missing value)

- **current_address_months_count** - Months at current address; ranges from [-1, 429] (-1 = missing value)
- **customer_age** - Age in years rounded down to decade; ranges from [10, 90] years
- **days_since_request** - Days since application was submitted; ranges from [0, 79] days
- **intended_balcon_amount** - Initial amount transferred; ranges from [-16, 114] (negatives = missing value)
- **zip_count_4w** - Applications in the same zip code in the last 4 weeks; ranges from [1, 6830]
- **velocity_6h** - Applications per hour in the last 6 hours; ranges from [-175, 16818]
- **velocity_24h** - Applications per hour in the last 24 hours; ranges from [1297, 9586]
- **velocity_4w** - Applications per hour in the last 4 weeks; ranges from [2825, 7020]
- **bank_branch_count_8w** - Applications at the bank branch in the last 8 weeks; ranges from [0, 2404]
- **date_of_birth_distinct_emails_4w** - Emails for applicants with the same birthdate in the last 4 weeks; ranges from [0, 39]
- **credit_risk_score** - Internal risk score; ranges from [-191, 389]
- **bank_months_count** - Age of the previous account in months; ranges from [-1, 32] (-1 = missing value)
- **proposed_credit_limit** - Proposed credit limit; ranges from [200, 2000]
- **session_length_in_minutes** - User session length on the banking website; ranges from [-1, 107] (-1 = missing value)
- **device_distinct_emails_8w** - Distinct emails associated with the device in the last 8 weeks; ranges from [-1, 2] (-1 = missing value)
- **device_fraud_count** - Number of fraudulent applications with used device.[0, 1]
- **month** - Month when the application was made. [0, 7]

Binary

- **email_is_free** - Whether email domain is free
- **phone_home_valid** - Validity of provided home phone number
- **phone_mobile_valid** - Validity of provided mobile phone number
- **has_other_cards** - Whether the applicant has other cards with the bank
- **foreign_request** - If the request origin country differs from the bank's country
- **keep_alive_session** - Whether the user enabled "remember me"

0.3 Methodolgy

0.3.1 Data Preprocessing

-{handling missing values} -{removing or inputing outliers} -{flagging inputed values}

-{Checking for missing values there is none but based on variable description some missing could be hideen due to -1 }

To increase the recall of the final model, the dataset needs to be as useful as possible. One part of this cleaning process is to remove any features that do not add anything to the model. One feature that does not add anything to the model is the device_fraud_count feature. Every single instance has the same value for that feature, 0, and therefore it cannot help detect fraud.

Another facet of the cleaning process is to deal with missing values. {deal with this later}

There are also some features in the dataset that use a negative number to encode a missing

value. Using these negative numbers directly would add bias in the model, so something needs to be done with them. We used two methods to deal with negative numbers in the dataset: Either remove the feature entirely from the model, or impute the feature.

To decide what features to remove, we calculated what percentage of each feature are missing values, and threw out the values with over 50% missing. As shown by the table below, the features with over 50% missing values are prev_address_months_count and intended_balcon_amount.

	Missing Data Count	Percentage Missing
prev_address_months_count	712920	71.2920
current_address_months_count	4254	0.4254
intended_balcon_amount	742523	74.2523
bank_months_count	253635	25.3635
session_length_in_minutes	2015	0.2015
device_distinct_emails_8w	359	0.0359

For all the other features with negative values, we used imputation to fill the values in. Imputation is the process of filling missing values with a reasonable placeholder value, so that the fitted model doesn't fit as poorly as it would with the original values in it. For this dataset, we used the median value in the feature for the placeholder. Finally, we also added a flag to any imputed columns. This was done to show our model that the features were imputed, so it could tell the difference between changed and unchanged data.

Another cleaning technique is one-hot encoding. One-hot encoding is a process used to separate categorical features into binary ones. For each option in the categorical feature, a new binary feature is added, which is turned on (a "1") when the categorical feature is on that option, and turned off (a 0) otherwise. This splits the categorical feature into binary features which shows what option the categorical feature is on, but without needing a categorical feature to show it.

For this dataset, we used one-hot encoding to turn the categorical features into features with numbers in them. This allows us to use models we would have been able to use otherwise, like tree-based models.

0.3.2 Exploratory Data Analysis

{Go over the summary of the Numeric EDA this is what i found so far}

current_address_months_count: right skewed customer age: centered around younger applicatns day_since_request: most applicat proces in 0.02 and mean of 1.02 days with some outliers velocity: Large range with negative values so significat outliers zip_count-4w: a wide range between regions credit_risk_score: wide range and negative (skewed) session length: wide range

income: Higher income correlates with fraud. name_email_similarity: Lower similarity suggests fraud. credit_risk_score: Higher scores correlate with fraud. proposed_credit_limit: Higher credit limits are indicative of fraud. current_address_months_count & customer_age: Older addresses and older applicants are associated with fraud.

Binned Data: Leverage Bin Groups in able to predict fraudlent behavior

Exploratory data analysis is the process of analyzing a dataset in order to find out its characteristics, especially if those characteristics help detect changes in the response. To accomplish our analysis, we used two methods: summary statistics, and binned data.

We calculated the following metrics for our summary statistics: mean, standard deviation, minimum, 25% quartile, 50% quartile, 75% quartile, maximum, mean within legit and fraudulent instances, and standard deviation within legit and fraudulent instances.

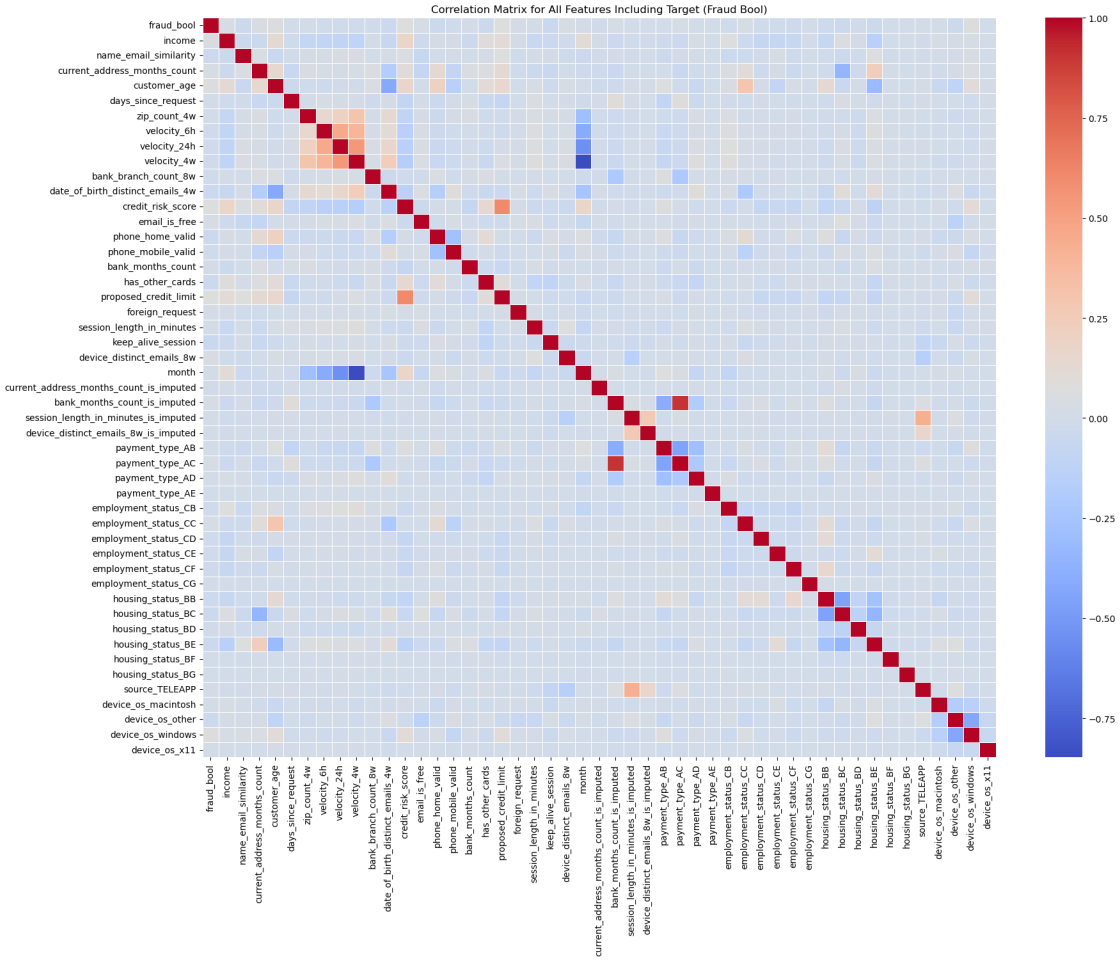
-{talk about what we found during catagorical}

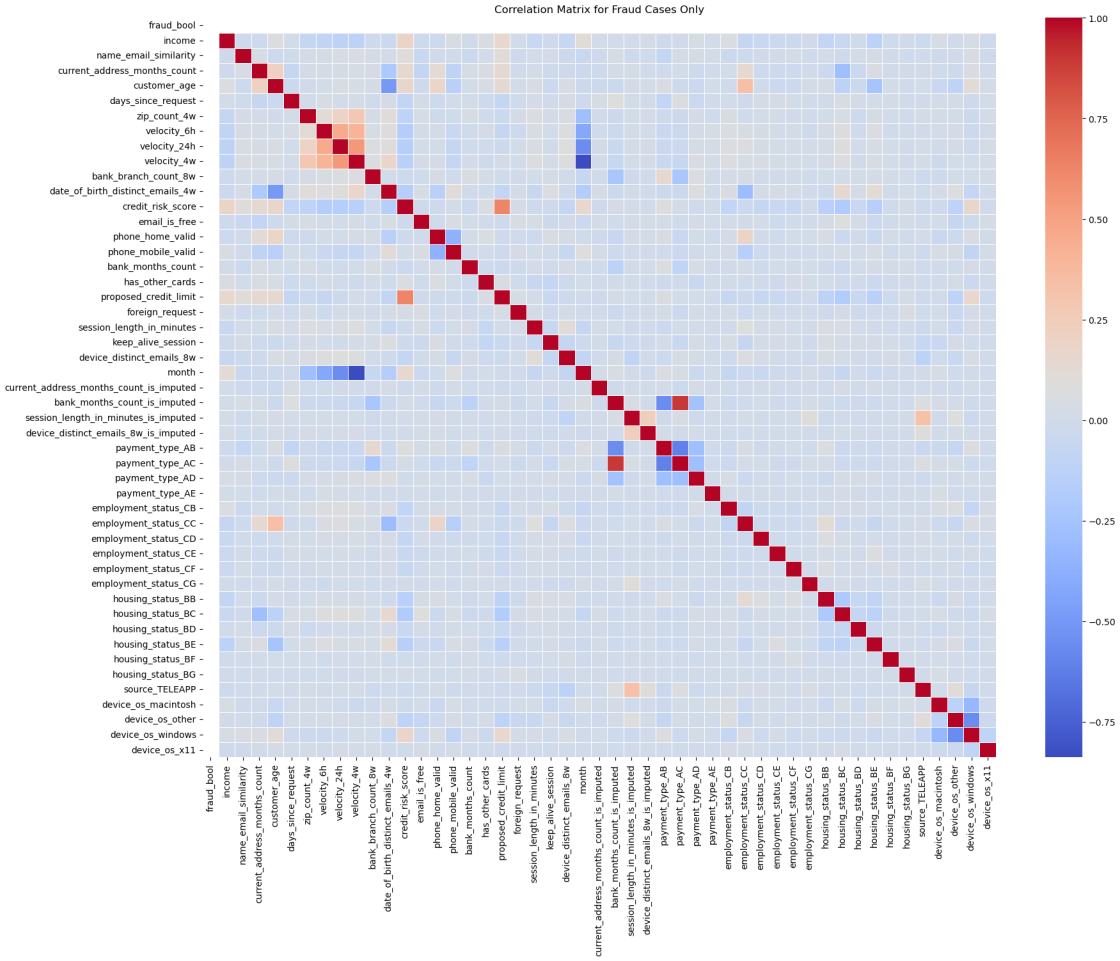
```
payment_type: AC Has the highest AE has the lowest
employment_status: CC has the highest, with CF and CE are lower
housing_status: BA has very high fraud, BE BG has much lower
source: Teleapp is higher
device_os: Windows has the highers, with others being much lower
```

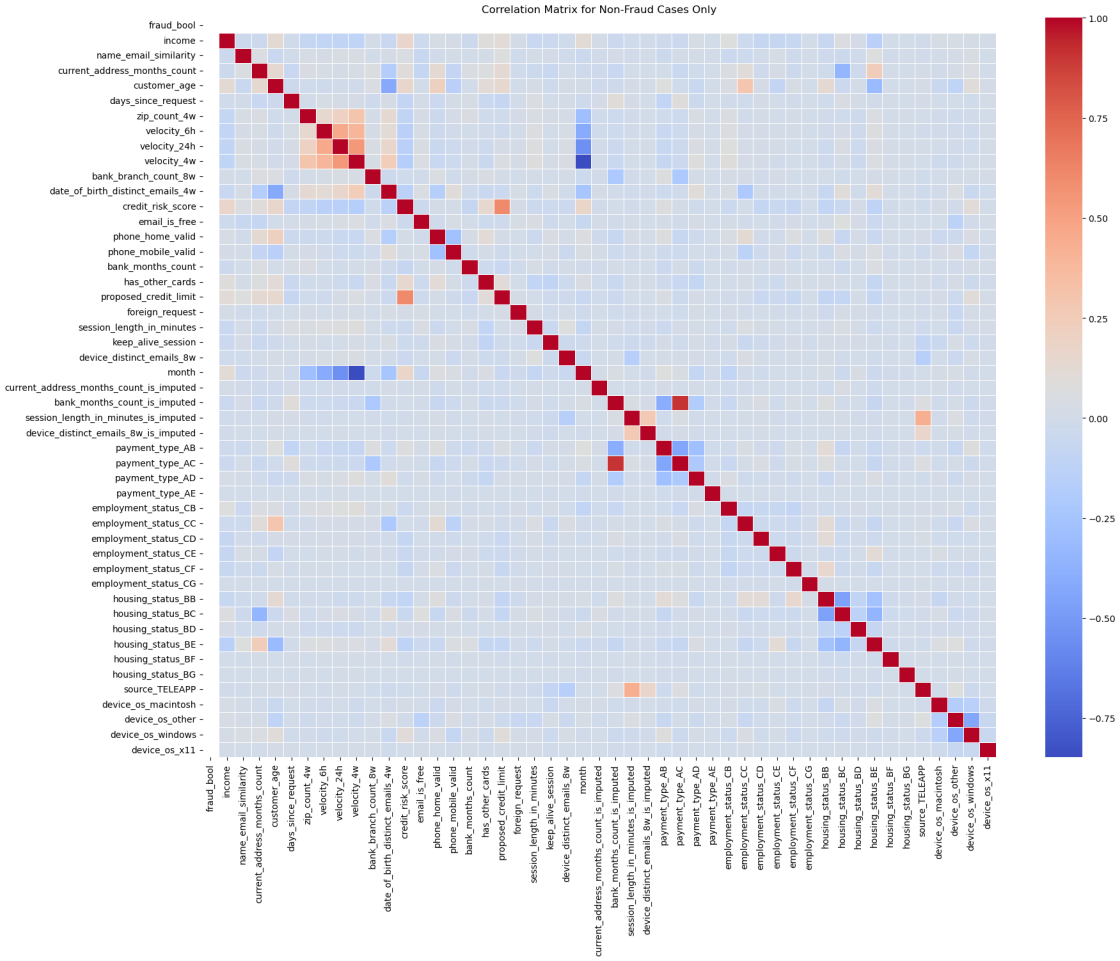
[31]:

	Feature	Count_True	Fraud_Percentage_True	Count_False	Fraud_Percentage_False
0	email_is_free	529886	1.375956	470114	0.795126
1	phone_home_valid	417077	0.669181	582923	1.413223
2	phone_mobile_valid	889676	1.054429	110324	1.493782
3	has_other_cards	222988	0.417511	777012	1.299594
4	foreign_request	25242	2.198716	974758	1.074523
5	keep_alive_session	576947	0.653093	423053	1.716333

-Free email domains, invalid phone numbers, and foreign requests show significantly higher fraud rates -applicants without other cards and non-remembered sessions are correlated with increased fraud


Correlation Matrix for All Features Including Target (Fraud Bool)
fraud_bool
income
name_email_similarity
current_address_months_count
customer_age
days_since_request
zip_count_4w
velocity_6h
velocity_24h
velocity_4w
bank_branch_count_8w
date_of_birth_distinct_emails_4w
credit_risk_score
email_is_free
phone_home_valid
phone_mobile_valid
bank_months_count
has_other_cards
proposed_credit_limit
foreign_request
session_length_in_minutes
keep_alive_session
device_distinct_emails_8w
month
current_address_months_count_is_imputed
bank_months_count_is_imputed
session_length_in_minutes_is_imputed
device_distinct_emails_8w_is_imputed
payment_type_AB
payment_type_AC
payment_type_AD
payment_type_AE
employment_status_CB
employment_status_CC
employment_status_CD
employment_status_CE
employment_status_CF
employment_status_CG
housing_status_BB
housing_status_BC
housing_status_BD
housing_status_BE
housing_status_BF
housing_status_BG
source_TELEAPP
device_os_macintosh
device_os_other
device_os_windows
device_os_x11
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Correlation Matrix for Fraud Cases Only
fraud_bool
income
name_email_similarity
current_address_months_count
customer_age
days_since_request
zip_count_4w
velocity_6h
velocity_24h
velocity_4w
bank_branch_count_8w
date_of_birth_distinct_emails_4w
credit_risk_score
email_is_free
phone_home_valid
phone_mobile_valid
bank_months_count
has_other_cards
proposed_credit_limit
foreign_request
session_length_in_minutes
keep_alive_session
device_distinct_emails_8w
month
current_address_months_count_is_imputed
bank_months_count_is_imputed
session_length_in_minutes_is_imputed
device_distinct_emails_8w_is_imputed
payment_type_AB
payment_type_AC
payment_type_AD
payment_type_AE
employment_status_CB
employment_status_CC
employment_status_CD
employment_status_CE
employment_status_CF
employment_status_CG
housing_status_BB
housing_status_BC
housing_status_BD
housing_status_BE
housing_status_BF
housing_status_BG
source_TELEAPP
device_os_macintosh
device_os_other
device_os_windows
device_os_x11
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75



0.3.3 Model Building