# Summary of prelimitary task | **Big Data** Boys

### Step 1

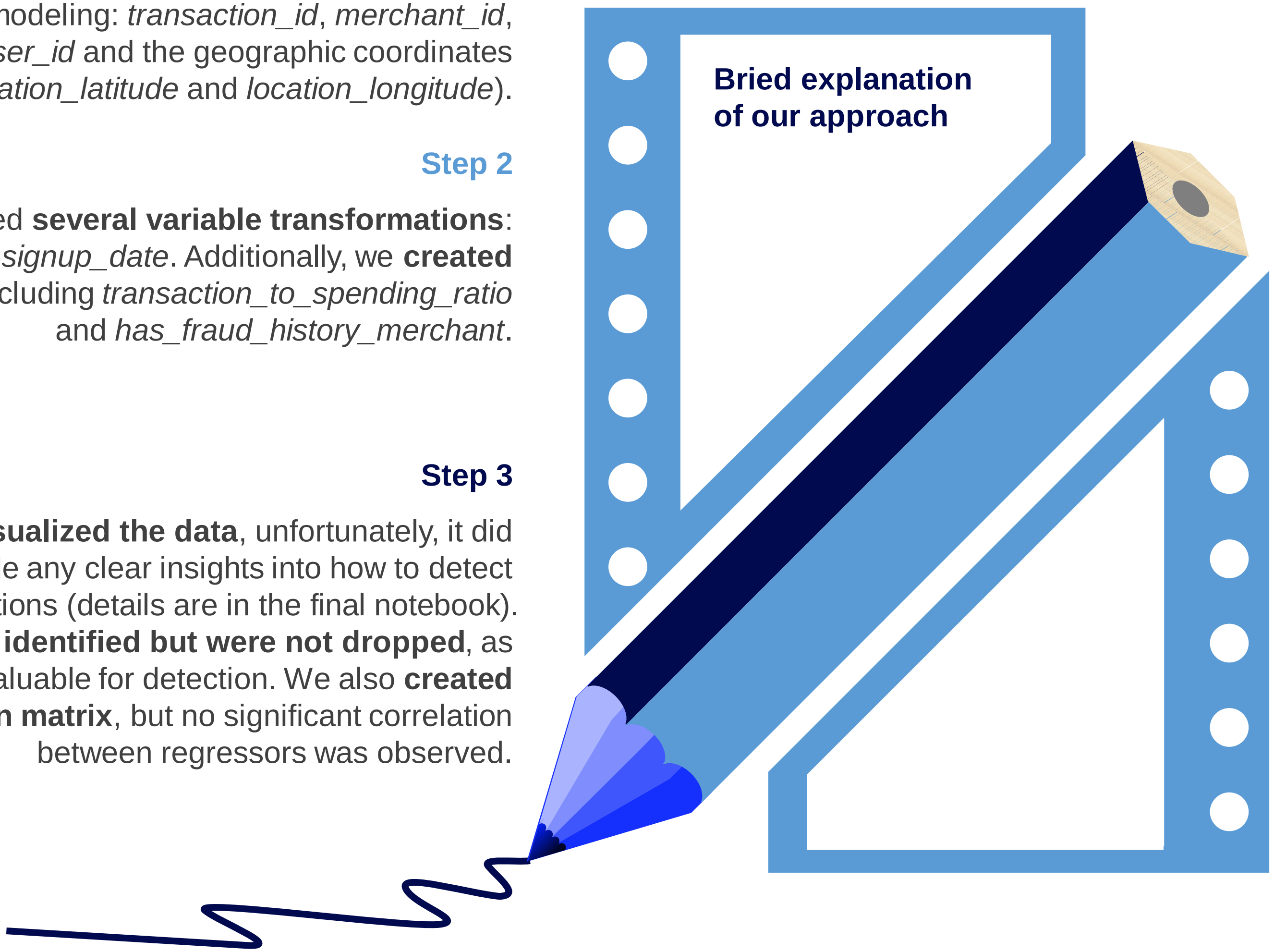
We have **deleted the key variables** that were not useful for modeling: *transaction_id*, *merchant_id*, *user_id* and the geographic coordinates (*location_latitude* and *location_longitude*).

### Step 2

We then performed **several variable transformations**: *timestamp* and *signup_date*. Additionally, we **created new variables**, including *transaction_to_spending_ratio* and *has_fraud_history_merchant*.

### Step 3

Although **we visualized the data**, unfortunately, it did not provide any clear insights into how to detect fraudulent transactions (details are in the final notebook). **Outliers were identified but were not dropped**, as they may still be valuable for detection. We also **created a correlation matrix**, but no significant correlation between regressors was observed.




### Step 4

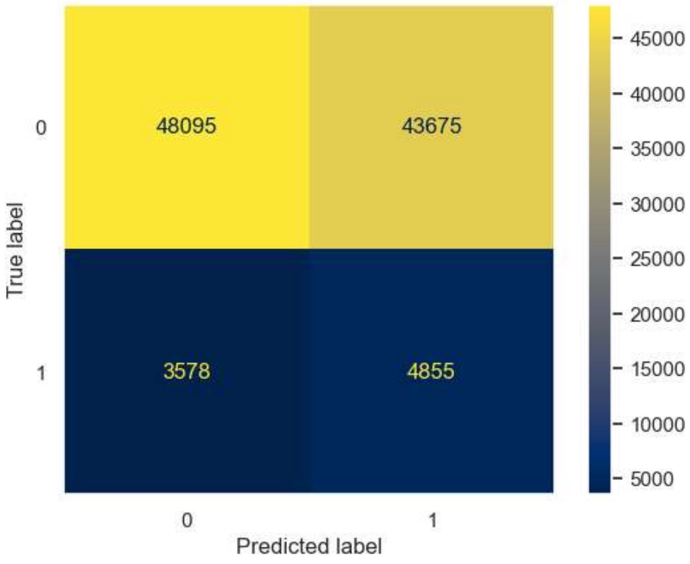
It was evident that **class imbalance was present in the data**, with 91.52% of the observations being non-fraudulent. To address this, we applied **oversampling**, **SMOTE** (Synthetic Minority Oversampling Technique), and **undersampling**. The best-fitting technique for this dataset was undersampling.

### Step 5

**After label encoding**, we proceeded to create models to predict our target variable. We **experimented with various combinations of training datasets and machine learning models with different hyperparameters**, including Random Forest, Decision Trees, SVC, XGBoost, and others. **The best model was XGBoost**, with slightly higher performance metrics.

# Summary of final models | **Big Data** Boys

| Metric | Value |
|---|---|
| Accuracy | 0.52842 |
| Precision | 0.10004 |
| MER | 0.47152 |
| Recall | 0.57571 |
| F1 Score | 0.17046 |
| Specificity | 0.52408 |
| AUC-ROC | 0.57150 |
| AUC-PR | 0.10412 |




| Feature | Importance |
|---|---|
| *has_fraud_history_merchant* | 0.185169 |
| *trust_score* | 0.073373 |
| *risk_score* | 0.049449 |
| *is_international* | 0.027050 |
| *session_length_seconds* | 0.003379 |
| *sum_of_monthly_installments* | 0.001855 |
| *account_age_months_user* | 0.001808 |
| *avg_transaction_amount* | 0.001070 |

## Explanation

Based on the confusion matrix and performance metrics, the **XGBoost model demonstrates highest effectiveness** in detecting fraudulent transactions among all models. Both **the AUC-ROC and AUC-PR scores indicate only moderate performance**. We've explained the situation more deeply in our jupyter notebook.

## Explanation

The SHAP analysis shows that the most influential factor in fraud prediction is *has_fraud_history_merchant*, followed by *trust_score*, *risk_score*, and *is_international*. Other features like *session_length_seconds*, *sum_of_monthly_installments*, and *avg_transaction_amount* have minor importance, while all other features show no impact.