Final project

PREDICTIVE ANALYSIS FOR FRAUD DETECTION

OUR TEAM

HYOJUNG LEE

MARIA GAYA

INTRODUCTION

PaySim:

a financial mobile money simulator devised for fraud detection.

As analysts, we can benefit from this simulated data to conduct fraud analytics overcoming the lack of public access to private records.

PROBLEM

1. The data contains fraudulent transactions
2. Only small proportion were flagged as fraud with poor accuracy

GOAL

Improve the financial fraud detection system
by identifying the best ML models

DATA ANALYTICS STEPS

COLLECT DATA

Data available
on Kaggle

CLEAN DATA

Exploration
Transformation

ANALYZE DATA

Modeling
Prediction

INTERPRET RESULTS

Evaluation

Classification

Classification for predicting binary target labels is a subcategory of supervised learning

- **Target**: Is it a fraud operation?
- **Features**: relevant columns

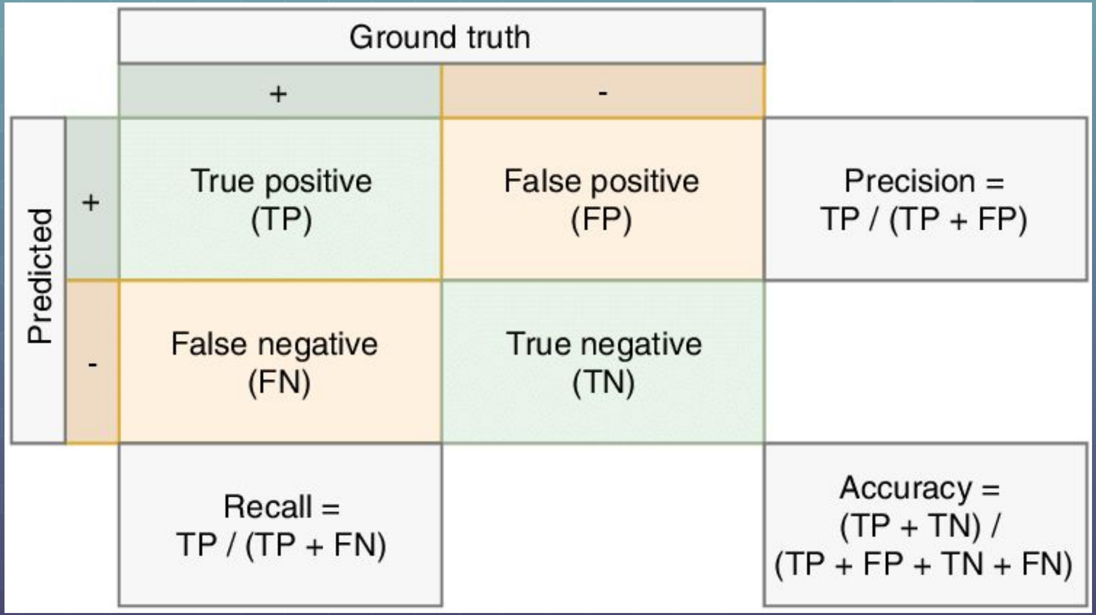
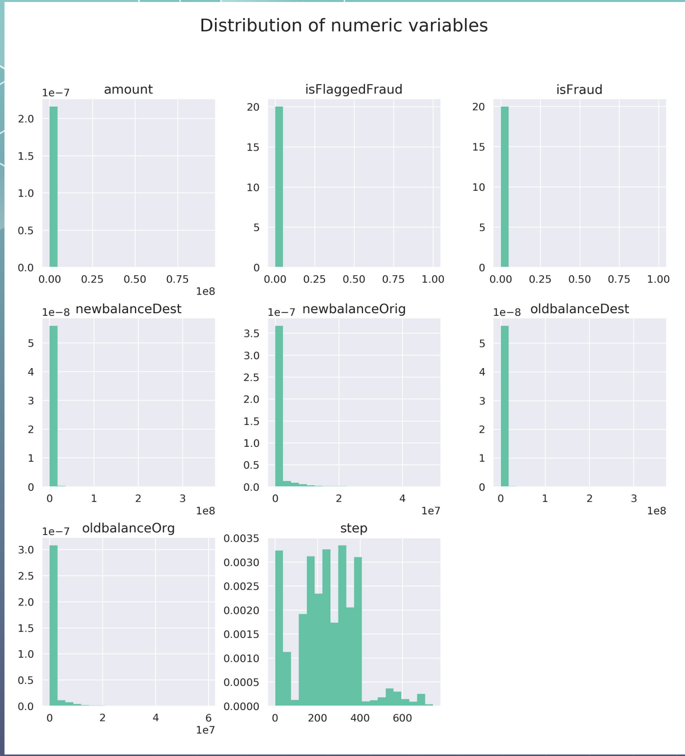


Figure from: https://www.researchgate.net/figure/Confusion-matrix-and-evaluation-metrics_fig3_334840641

FIRST SIGHT ANALYSIS I



Histograms generated by its counts normalized to form a probability density. The area under each histogram sums to 1.

6.362.620

Observations/Rows

11

Features/Columns

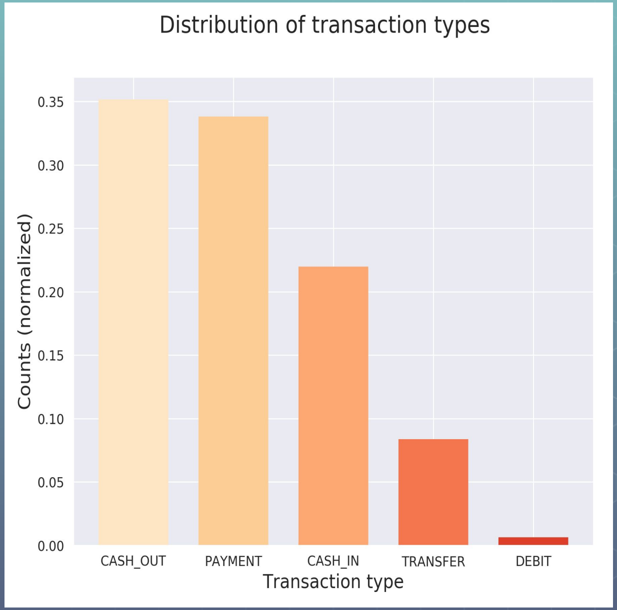
0

Missing values

8219

Fraudulent transactions CtoC

FIRST SIGHT ANALYSIS II



In order of frequency:
CASH_OUT (35 %) > PAYMENT (34 %) > CASH_IN (22 %) > TRANSFER (8 %) > DEBIT (1 %)

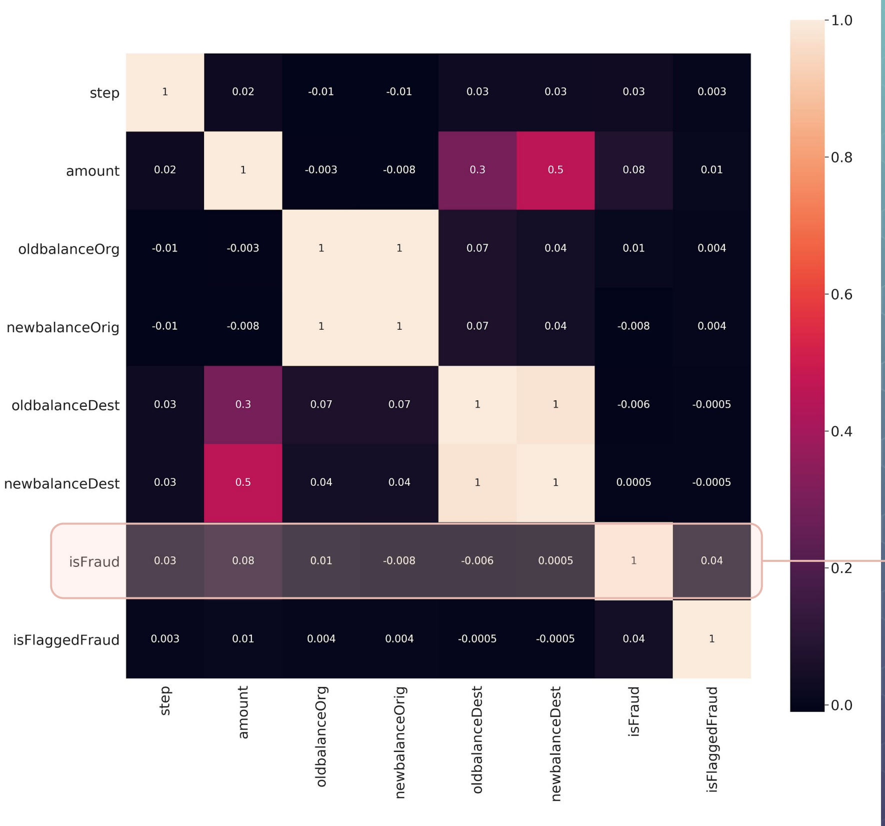
Valid (not-fraud):
99% (6,354,407 cases)

Fraud:
0.1 % (8,213 cases)

Flagged as Fraud:
0.0003 % (16 cases)

Frauds failed to flag:
8197 cases !

FIRST SIGHT ANALYSIS III



COLUMN NAME	r
amount	0.077
isFlaggedFraud	0.044
step	0.032
oldbalanceOrg	0.010
newbalanceDest	0.001
oldbalanceDest	-0.006
newbalanceOrig	-0.008

Low linear correlation with the target

Fraud vs. Valid

FIRST SIGHT ANALYSIS IV

A. Amount of each transaction

- ★ **Fraud**: 0 - 1.0 million (in local currency)
- ★ **Valid**: mostly < 0.2 million (in local currency)

B. Type of transaction

- ★ **Fraud**: Transfer, Cash out
- ★ **Valid**: Cash_out, Payment, Cash_in, Transfer, Debit

C. Operation time

- ★ **Fraud**: Generally active over the entire month
- ★ **Valid**: Less active after the first two weeks of the month

D. Client – Recipient type of each transaction

CC: Customer - Customer

CB: Customer - Business

BC: Business - Customer

BB: Business - Business

MODELING

LOGISTIC REGRESSION

NEURAL NETWORK

RANDOM FOREST

XGBOOST

EVALUATION METRICS

Confusion matrix

Table that is used to describe the performance of a classification model

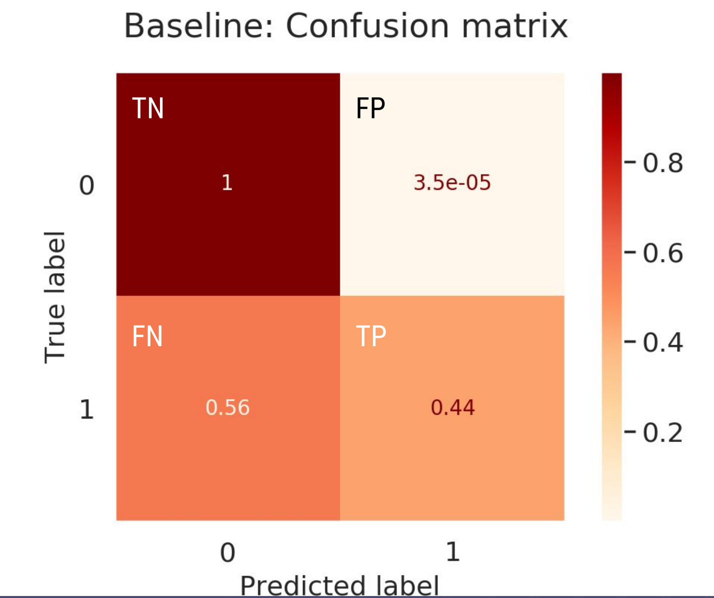
Area under curve

Curve that shows the tradeoff between precision and recall for different threshold

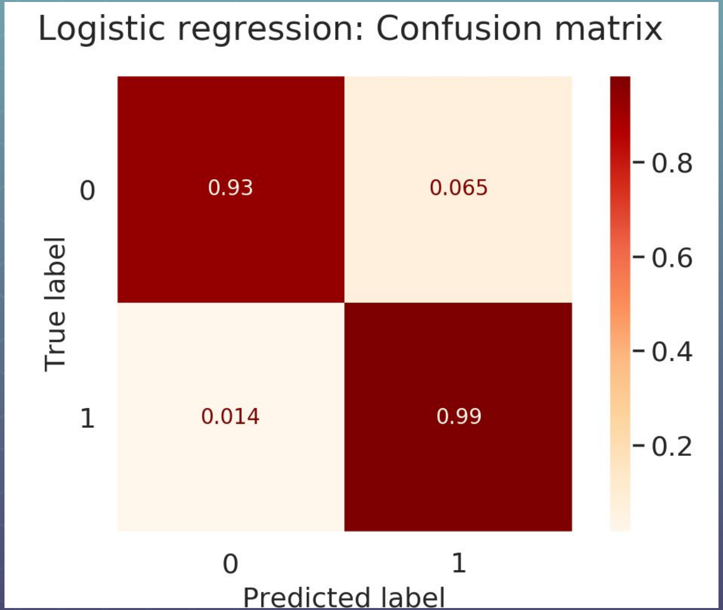
Recall rate

Out of all the positive classes, how much we predicted correctly. It should be high as possible.

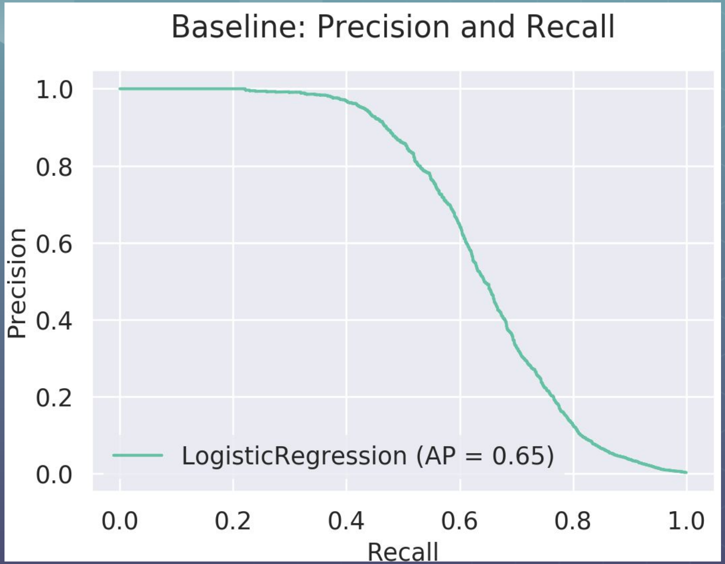
Logistic Regression



+Class weights

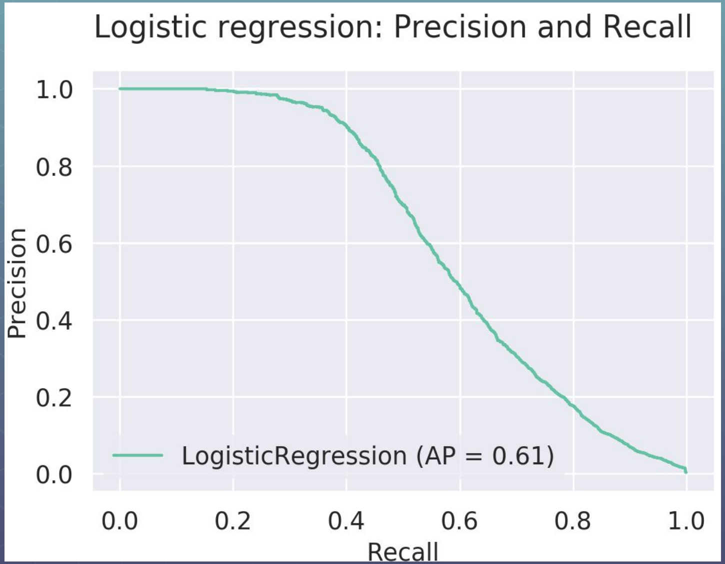


Logistic Regression



Recall score: 0.438

+Class weights



Recall score: 0.986

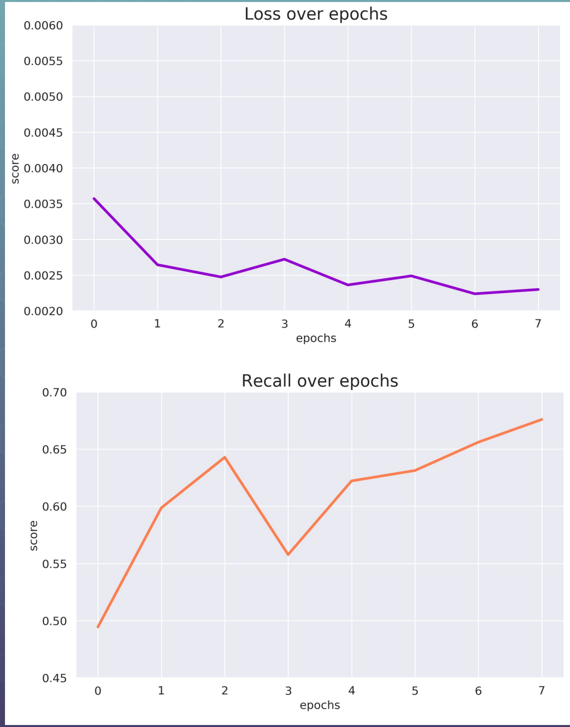
Neural Network

Sequential model

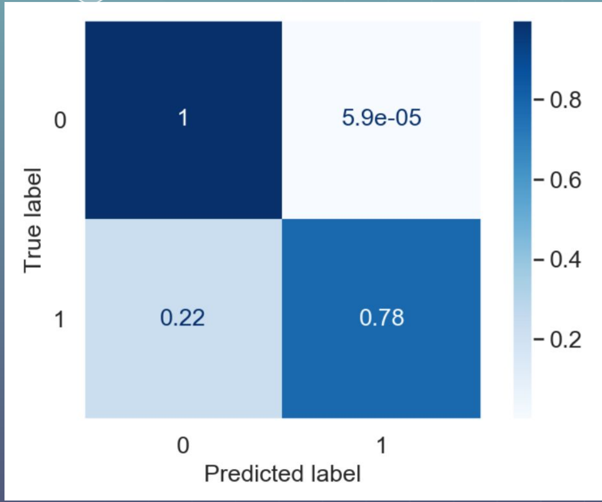
- Input: 10 features (4453834 samples)
- # Dense layers: 3-5
- # neurons: (32-24)-16-8-1
- # epochs: 3-8

Trial 1

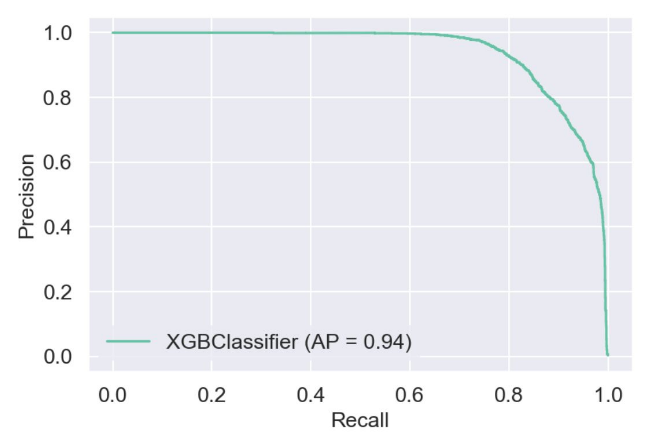
Trial 2



XGBoost



High rate of TP and FN



Recall score: 0.78

RESULTS – RECALL SCORES

0.985

LOGISTIC REGRESSION

XGBOOST

0.78

0.676

0.45

Random forest

Neural network

FINAL STATEMENTS

Conclusion

We detected fraudulent transactions with high precision and high recall by building highly functioning predictive classifier models

Future directions

1. Optimize the present models
2. Predict on more complicated data

Questions?

Our notebooks: https://github.com/CodeOp-tech/projectfraud-paysim

Hyojung Lee & Maria Gayà

www.codeop.com

THANKS