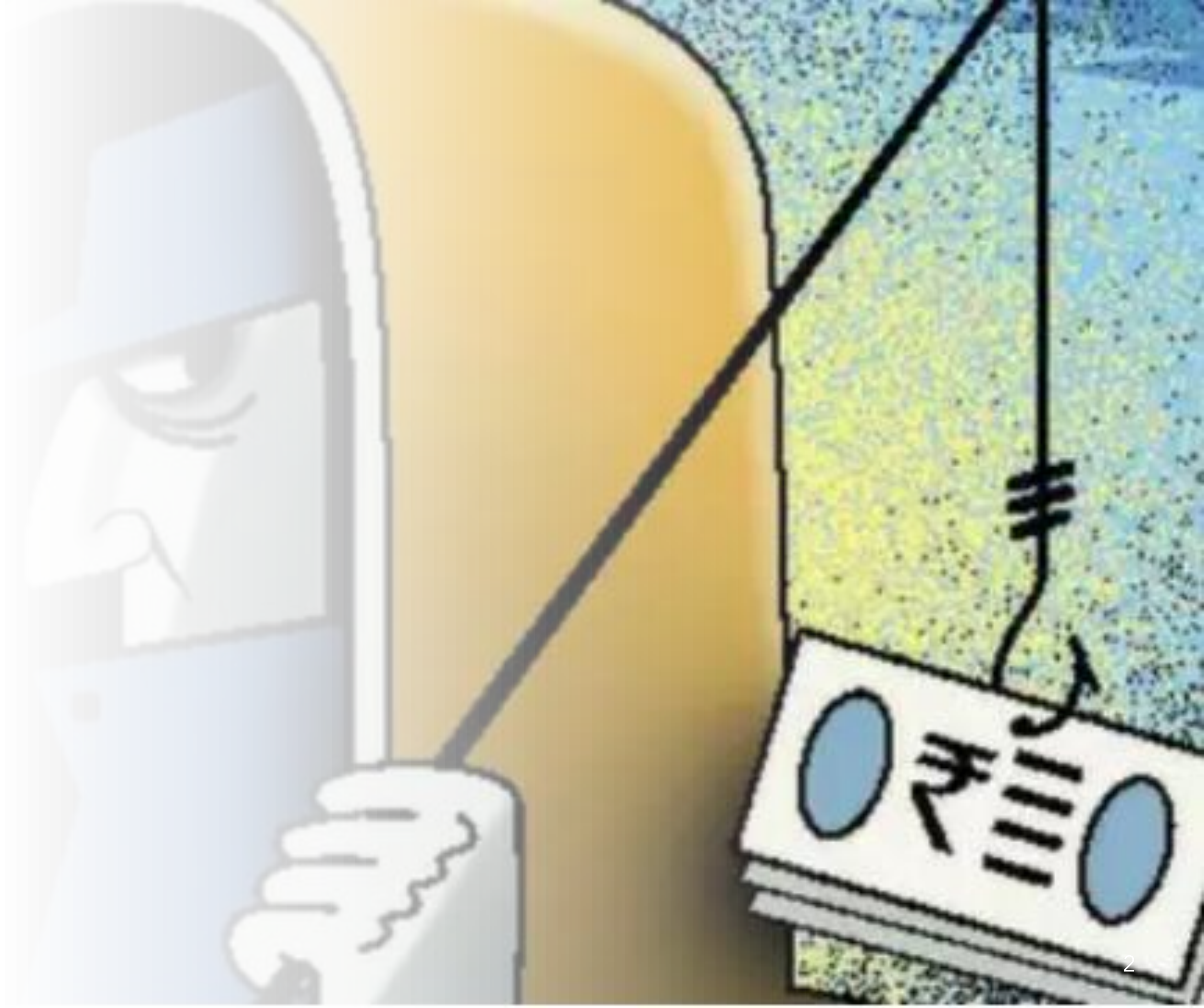
Alert Machines:
Detecting Fraud Before It Happens

Harry Dzeba

Introduction

- Credit card data for half a million people is currently on sale on the dark web
- Retailers worldwide will lose $130B over 5 years to credit card fraud, as people are making more transactions online
- 75% of all card fraud is on online transactions



Data

- 2019 Kaggle dataset by Vesta Corporation, a payment services company
- Data is anonymized, to protect the privacy of individuals involved
- Columns come in groups (C1-C14, V1-V339, M1-M9). The meaning of groups is known, but the exact natures of (most) individual columns is obfuscated
- The target value is isFraud binary column (1 means the transaction was fraudulent)

Objectives

1. Develop a machine-learning model that prevents fraudulent transactions while minimizing business disruption from flagging legitimate transactions

2. Make the model accurate without being reliant on sheer computing power

Agenda

Dealing with large dataset

Exploratory data analysis

Models and conclusions

1. Reduce memory usage

```
# merged dataset is huge - reduce memory footprint
reduce_mem_usage(train)
reduce_mem_usage(test)
```

executed in 2m 8s, finished 18:29:49 2022-02-22

```
Memory usage of dataframe is 1959.88 MB
Memory usage of dataframe is 1959.88 MB --> 533.44 MB (Decreased by 72.8%)
Memory usage of dataframe is 1677.73 MB
Memory usage of dataframe is 1677.73 MB --> 464.96 MB (Decreased by 72.3%)
```

How the function works:

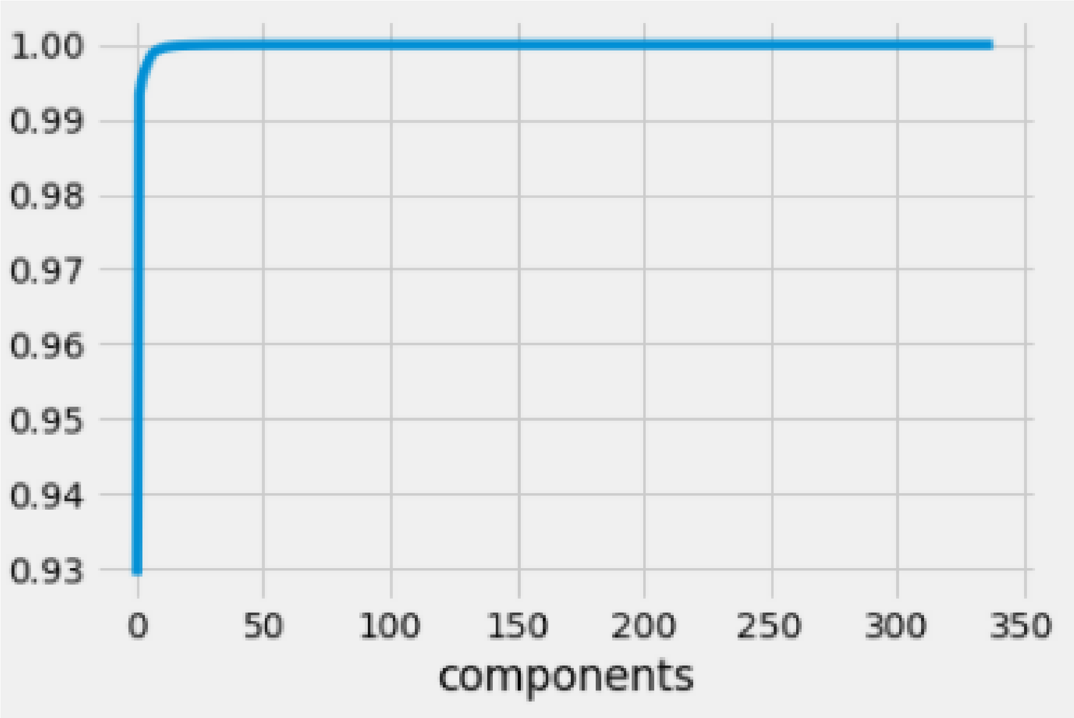
- it finds the min and the max value for each column
- assigns smallest memory-usage type to each columns, while making sure both min and max are still in range
- Int8 uses twice as little memory as Int16

2. PCA - Principal Component Analysis

- Used for dimensionality reduction (1 column = 1 dimension)
- Projects each data point onto only the first few principal components
- Tries to preserve as much of the data's variance as possible (our goal – 99%)

PCA – Put Columns Away

PCA for V group of columns (V1-V339)




```
number of  PCA features needed for 99% explained variance: 2
```

We can collapse 339 V-columns into only 2 while keeping 99% of explained variance

If PCA is so great, why not do it with other columns?

Loss of interpretability, as PCA features are linear combinations of other features	V columns are a group of obscure, synthetic columns, so there was no interpretability to begin with

3. Hash Encoding

- Label encoding assigns random integers to categorical values, which doesn't make sense when the values aren't ordered
- One Hot encoding creates new column for each unique categorical value
- Neither would work on this dataset
- Hash encoding did work - it maps each category in a feature to an integer within a predetermined range

Label Encoding

Food Name	Categorical #
Apple	1
Chicken	2
Broccoli	3

One Hot Encoding

Apple	Chicken	Broccoli
1	0	0
0	1	0
0	0	1

Hash encoding explained

How it works:

- Fixed number of new integer columns are created (usually 8 or 32) using cryptographic hash function
- Example: 3 columns with 50 unique values each: dummy encoding results in 150 new columns, vs 8 (or 32) for hash encoding
- Downside is 'collisions' – multiple values get hashed into the same integer

- ❖ Ostrich Algorithm – pretend the problem doesn't exist
- ❖ Even on 50% colliding features the loss is less than 0.5%

From Wikipedia, the free encyclopedia

In computer science, the **ostrich algorithm** is a strategy of ignoring potential problems on the basis that they may be exceedingly rare.

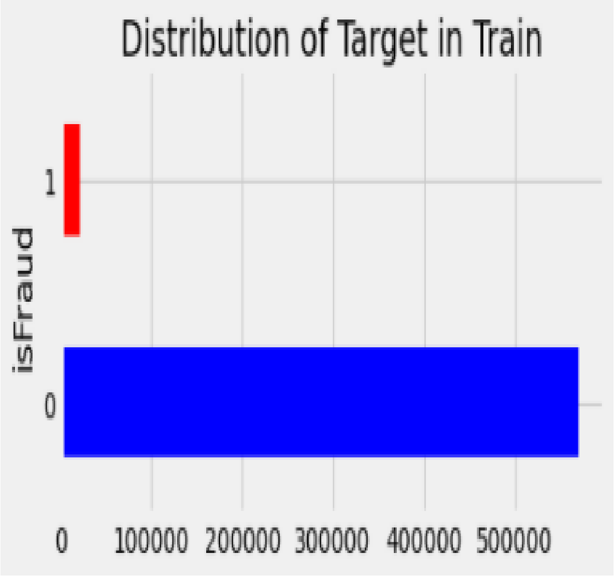
```
executed in 4h 2m 29s, finished 00:38:10 2022-02-23

Fold 1 started at Tue Feb 22 20:35:41 2022
Fold 2 started at Tue Feb 22 21:15:55 2022
Fold 3 started at Tue Feb 22 22:30:56 2022
Fold 4 started at Tue Feb 22 23:09:37 2022
Fold 5 started at Tue Feb 22 23:57:11 2022
CV mean score: 0.9237, std: 0.0153.
```

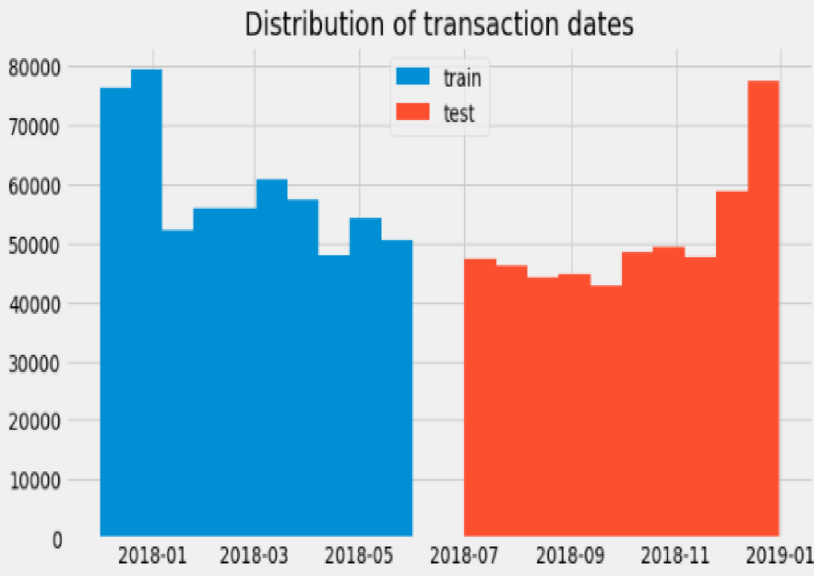
4. Custom K-fold validation for models

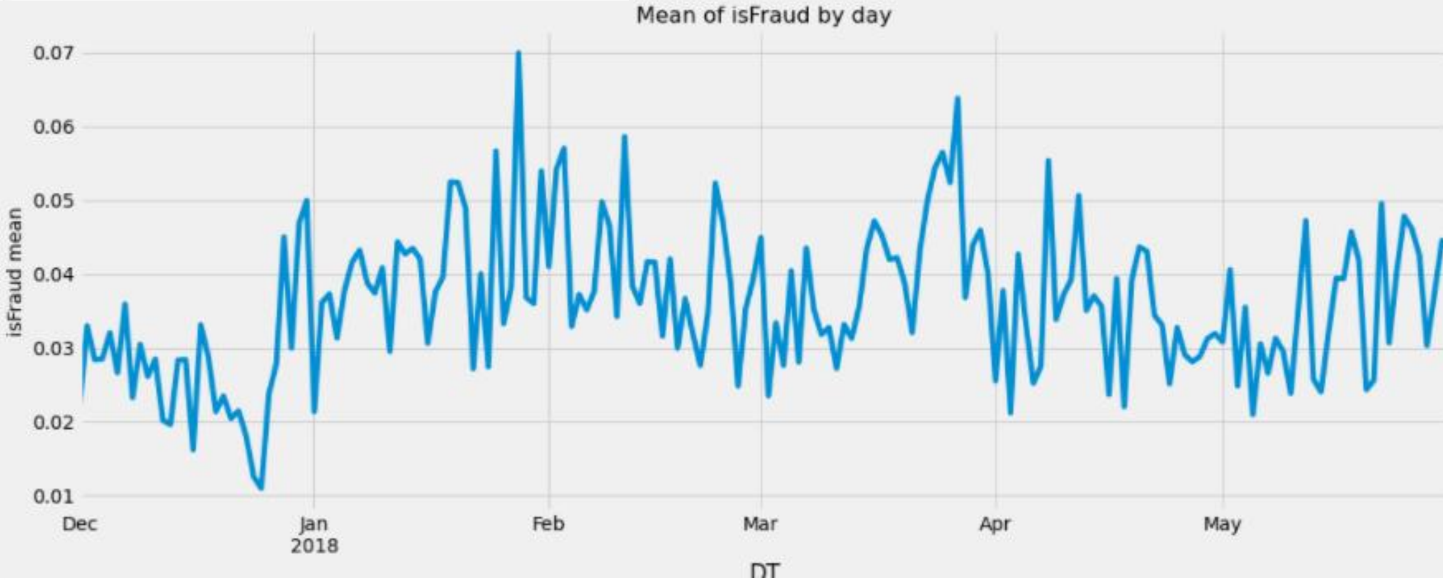
- Validating a model across 5 different test folds can take a long time. By writing a custom function that updates us after each fold begins, we can have a good estimate of how long it will take to run a model

EDA

- Fraud is found in only 3.5% of the rows
- Test and train data is chronological separated by one month
- Fraud rates drop during Christmas time

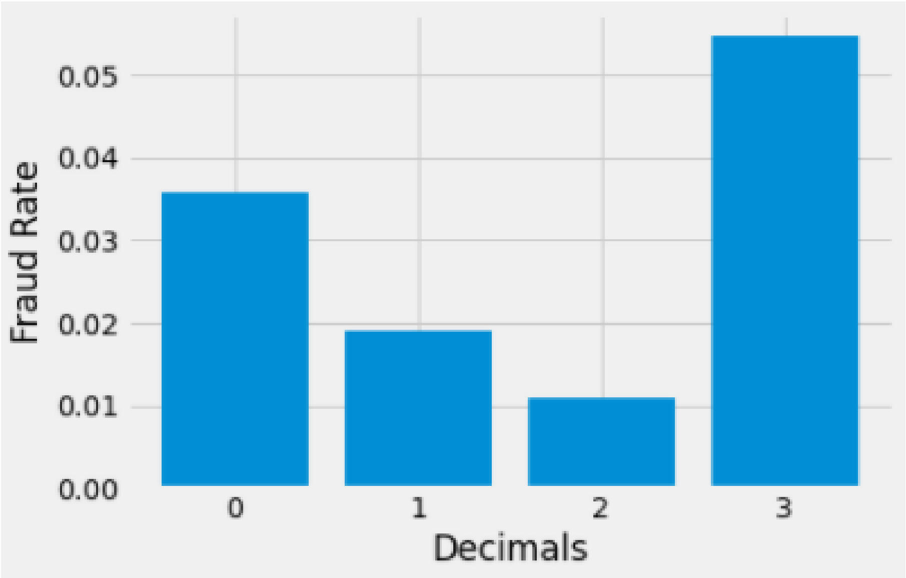


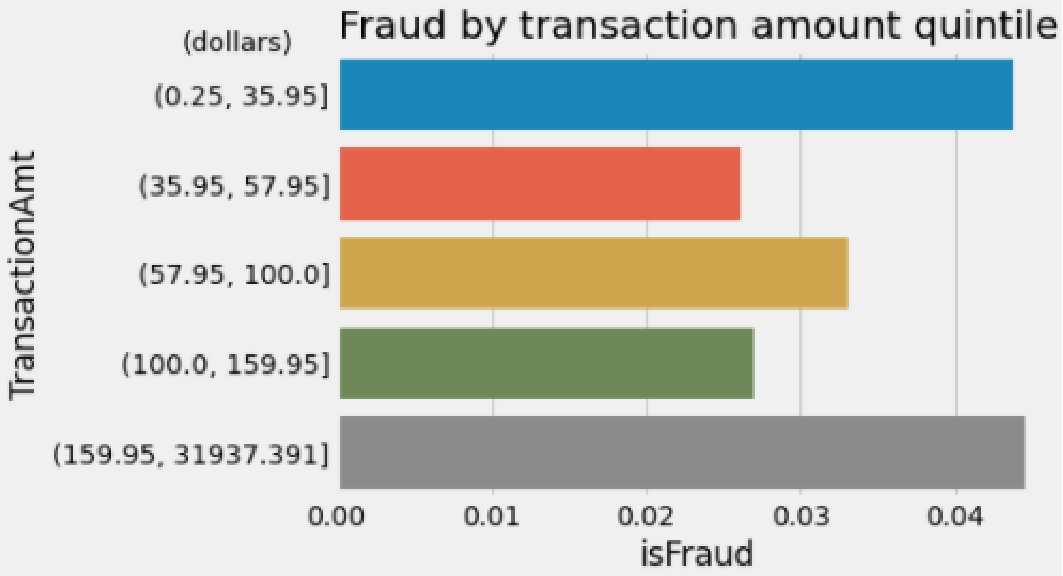




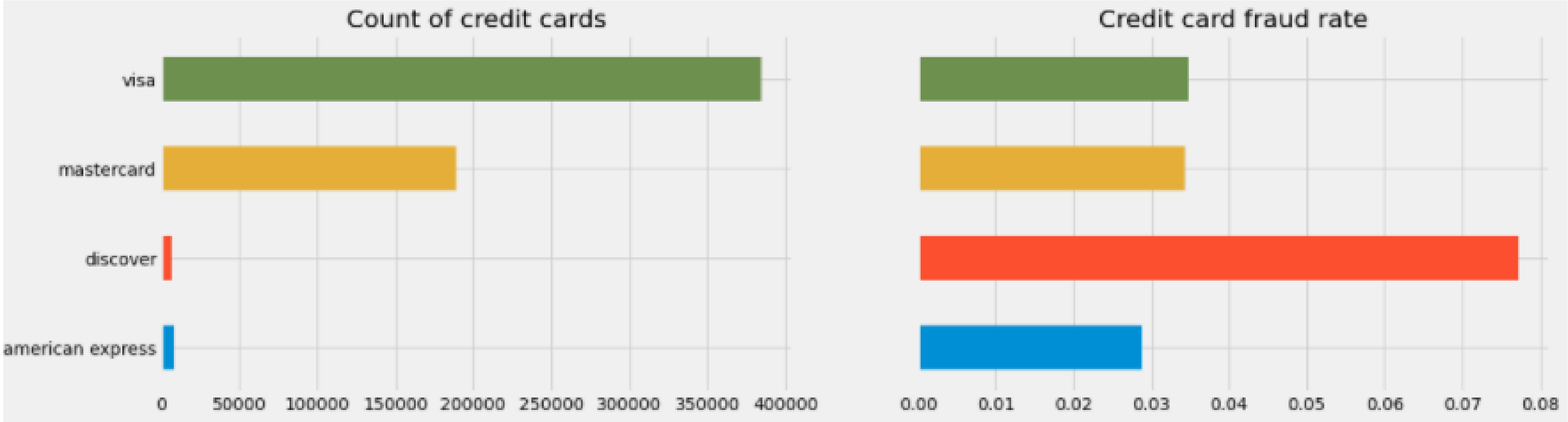
Decimals and Transaction Amounts

- International fraud – detected through number of decimals
- Fraud rates higher for very small and very large amounts



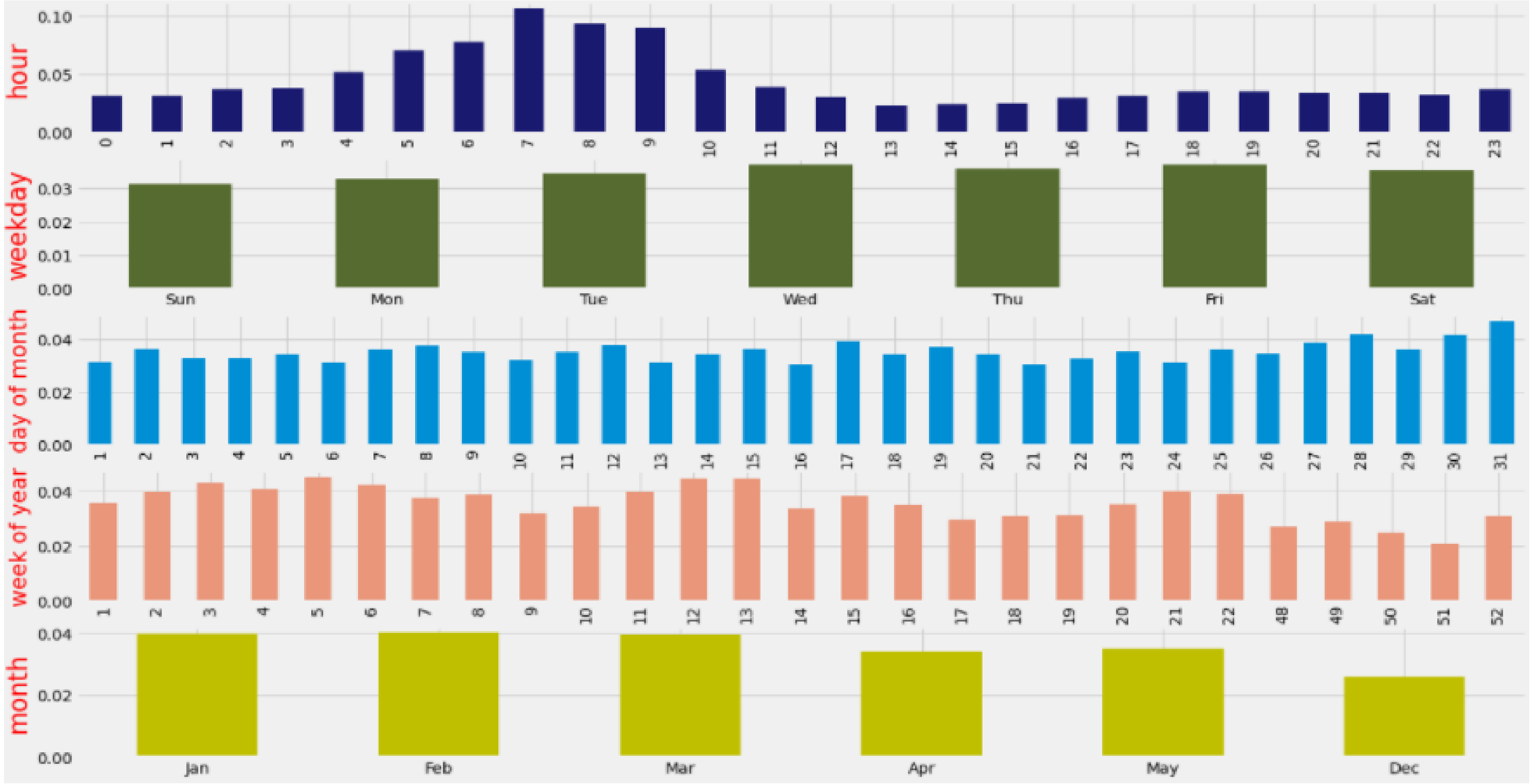


Discover – that you've been a victim of a fraud



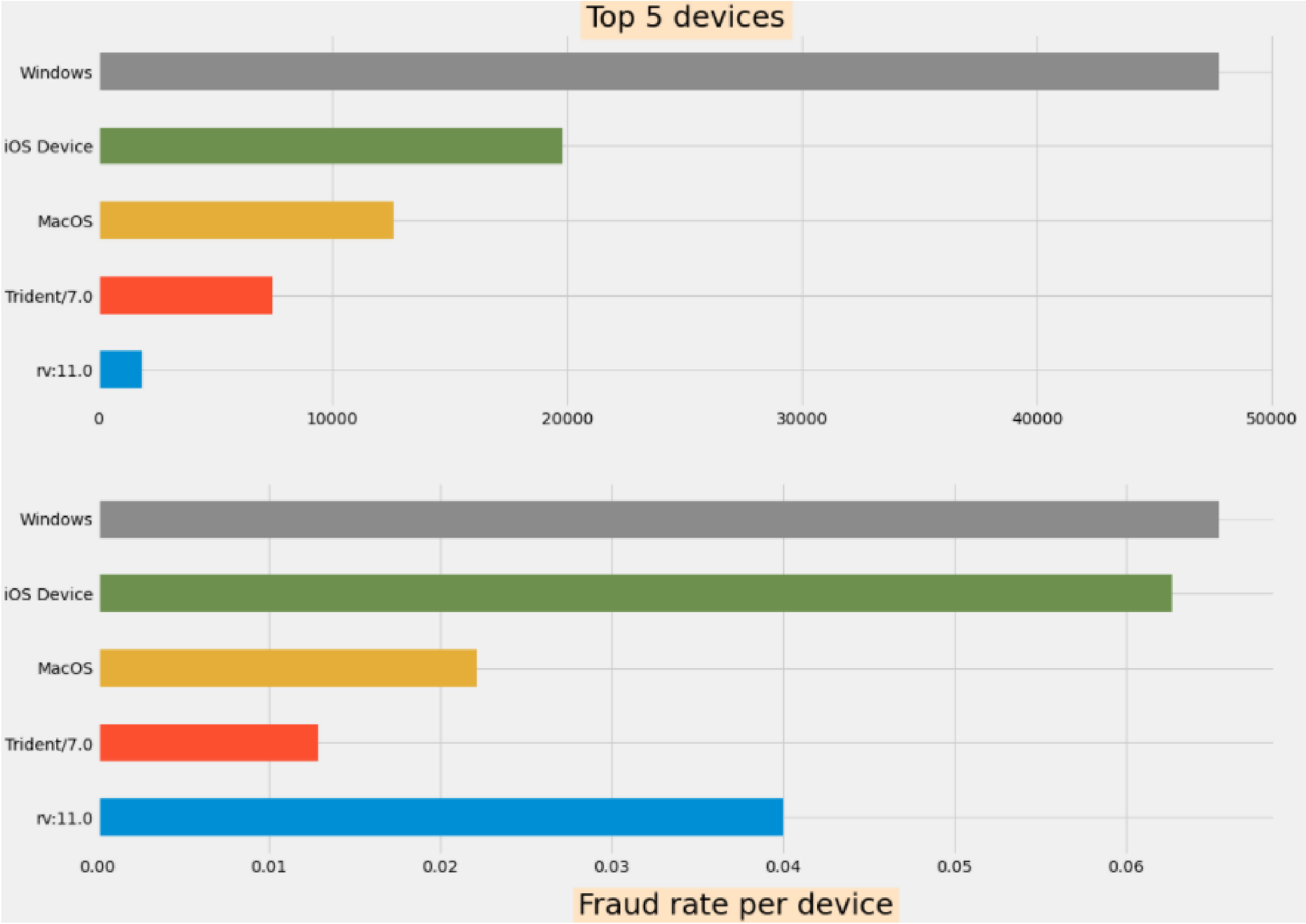
Date/time

New date/time columns created as fraud rate depends on time/day/week/month



Operating System

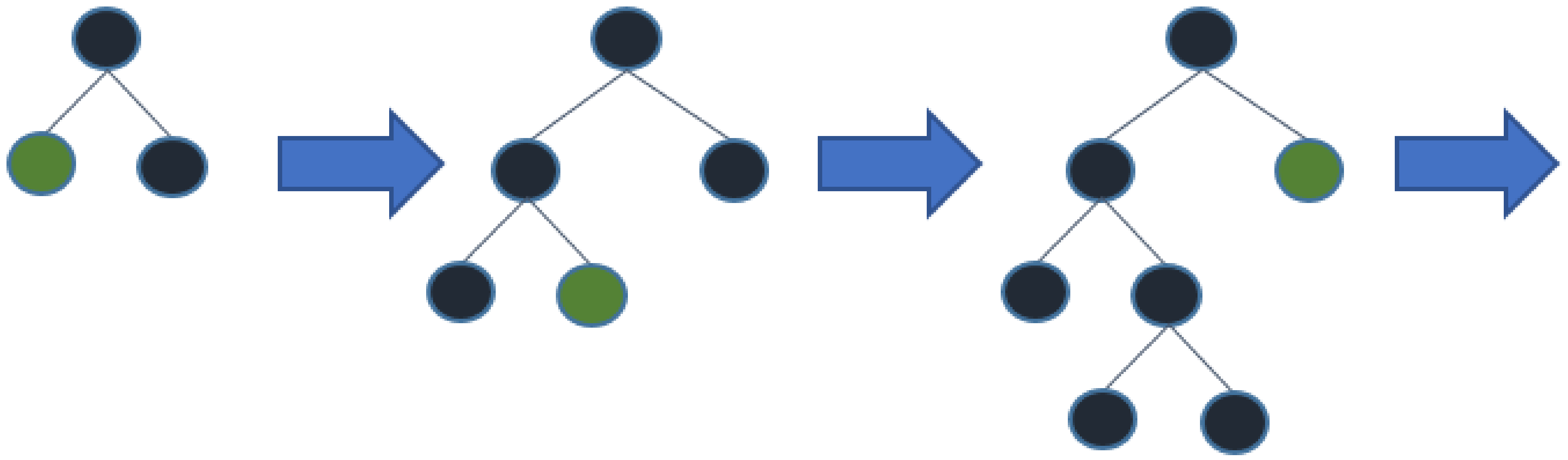
- Two most popular operating systems (iOS and Windows) are also the two with the highest fraud rates



Models

LightGBM – a decision tree algorithm focused on speed and efficiency

- Leaf-wise growth
- Histogram binning
- Internal handling of categorical features and null values



Leaf-wise tree growth

Models

(Different versions of LGBM, Random Forest, Neural Network)

	Model	CV mean ROC-AUC	ROC-AUC-train	ROC-AUC-test	Precision	Recall
0	LGBM1	0.924	0.925	0.888	0.869	0.287
1	LGBM2	0.914	0.912	0.872	0.864	0.269
2	LGBM3-wght	0.901	0.895	0.896	0.500	0.499
3	LGBM_enc1	0.924	0.925	0.899	0.812	0.320
4	LGBM_enc2_W	0.906	0.904	0.874	0.368	0.513
5	LGBM_enc32	0.924	0.925	0.900	0.819	0.320
6	RF1	0.897	0.976	0.887	0.843	0.259
7	CNN2-rlr	NaN	0.877	0.820	0.121	0.676

(before encoding)

(after encoding)

Tuning the model – 3 types of parameters

Tree Structure

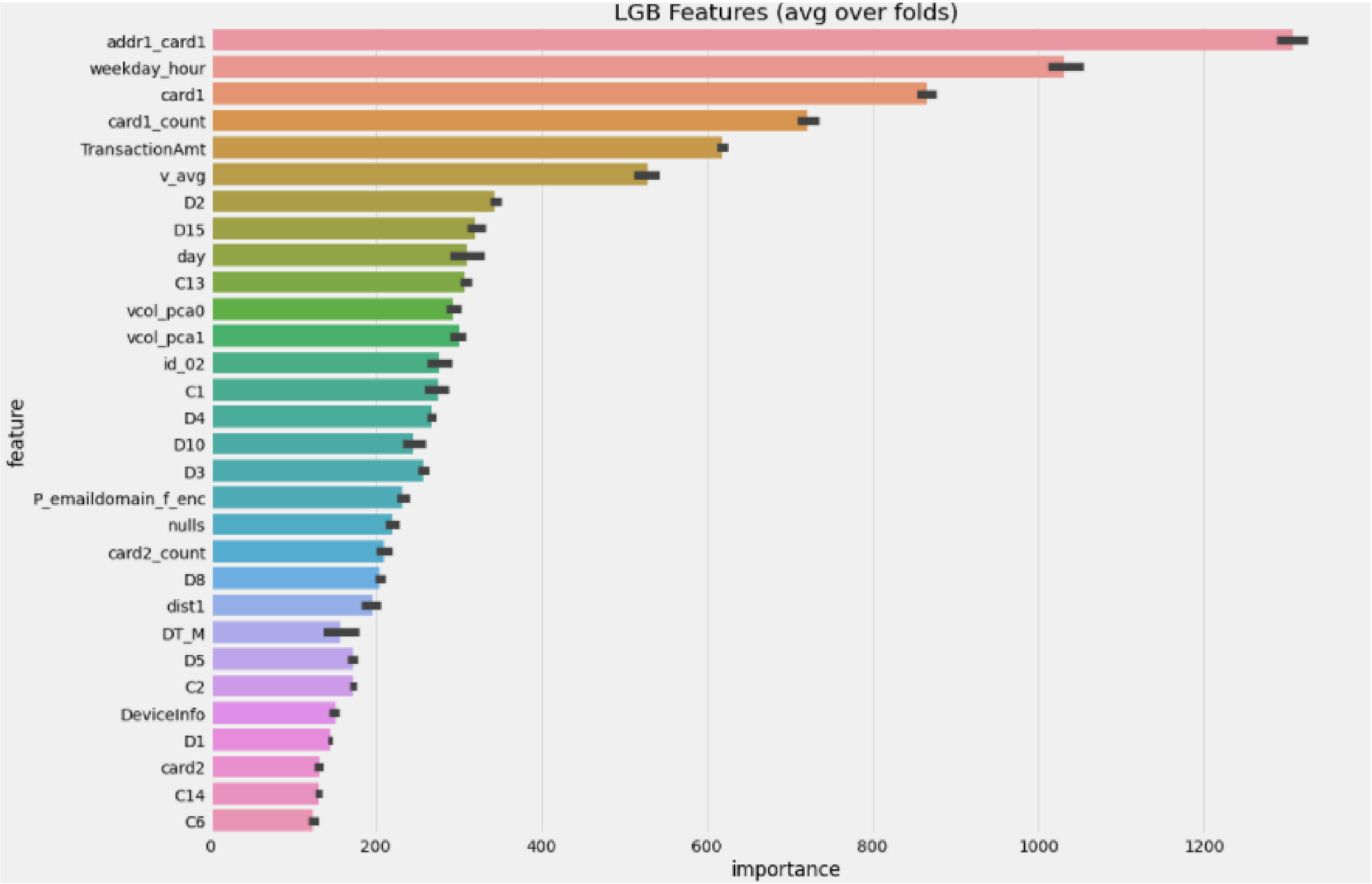
- num_leaves
- max_depth
- min_data_in_leaf

Accuracy

- n_estimators
- learning_rate

Combat overfitting

- reg_alpha
- feature_fraction
- min_gain_to_split



Conclusions: which features helped the most?

Next Steps

Present:

- We got AUC score of 0.9 despite reducing the dataset by 75%
- What could the score be if we had unlimited computing resources?

Future:

- Work on real, unprocessed credit card fraud data from a bank
- Broaden the scope of the war on fraud – use behavioral analytics to identify bad actors across a range of online misconducts such as illegal content, fake reviews, malware

Thanks for listening!