Credit Card Fraud Detection

Team Members:
Dingying (Daniel) Zhang
Ojaswa Garg
Vijay Krishna Nallapaneni
Zhaochen (Sam) Ye

Executive Summary

Business Problem and Background

Consumers in US have experienced a rising number of credit card frauds, which cost both consumers and merchants' time and money. The number of victims continues to rise and the amount of money affected is estimated to reach $30 billion this year

Data science is widely adopted across industries to facilitate data-driven decision making and solve troublesome tasks which were nearly impossible. Data science can help the finance industry by allowing automatic detection of fraudulent transactions

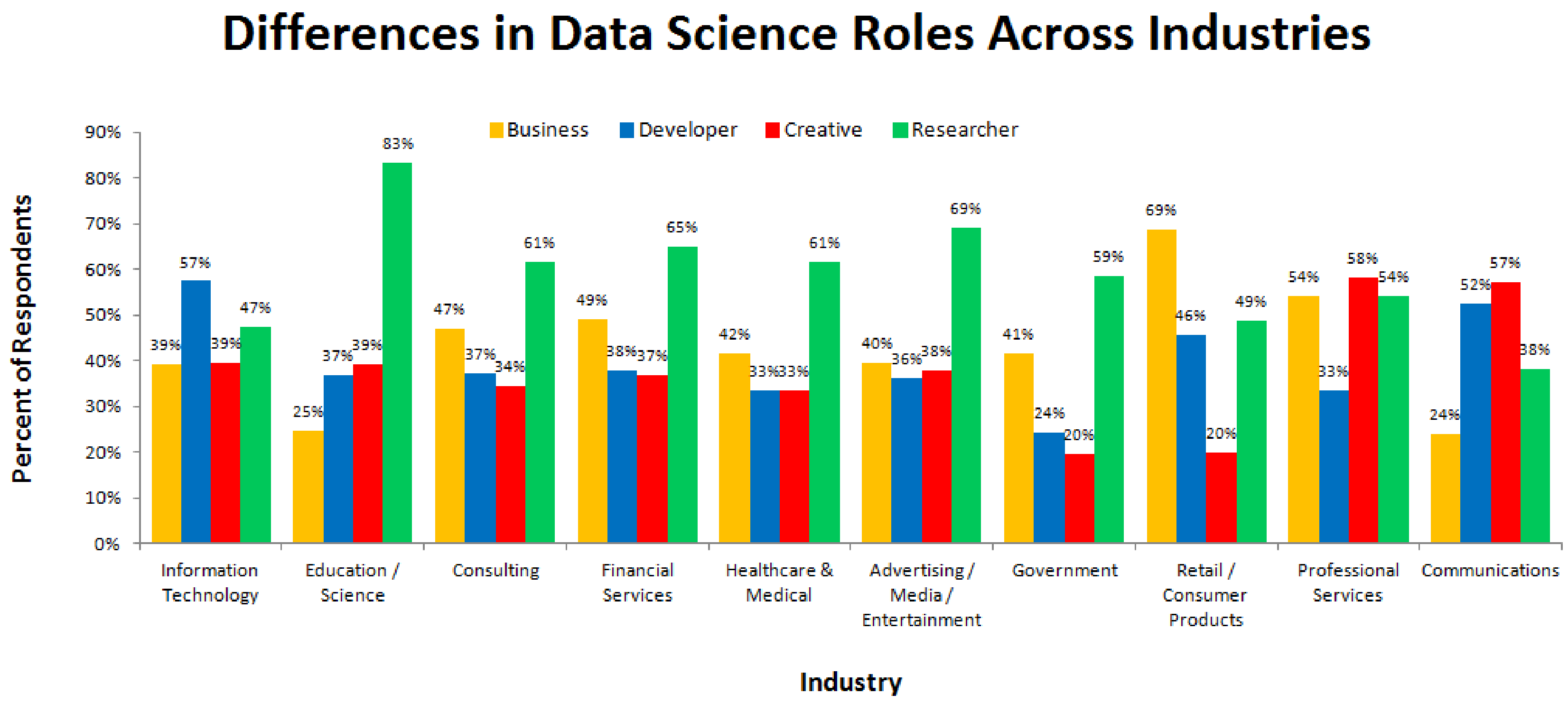
Methodology and Result

	Logistic Regression	Random Forest	Neural Network
Description	Starting Point to explore the model performance and allow key stakeholders to understand impacts of features	Ensemble tree-based model to avoid overfitting	Effort to capture impact of each feature as much as possible
Result	High Accuracy & High Recall Rate, but room for improvement with more relevant data sources		

Background

Data Science in Different Industries

There are many areas in finance industry that data science can create value



Data scientists play different roles across industries. Some industries leverage data science to facilitate business operations (e.g. retail / consumer products), a lot others focus on research (e.g. education and consulting)

Data Science in Financial Services Industry

Risk Management

Fraud Detection

Pricing Automation

Customer Experience

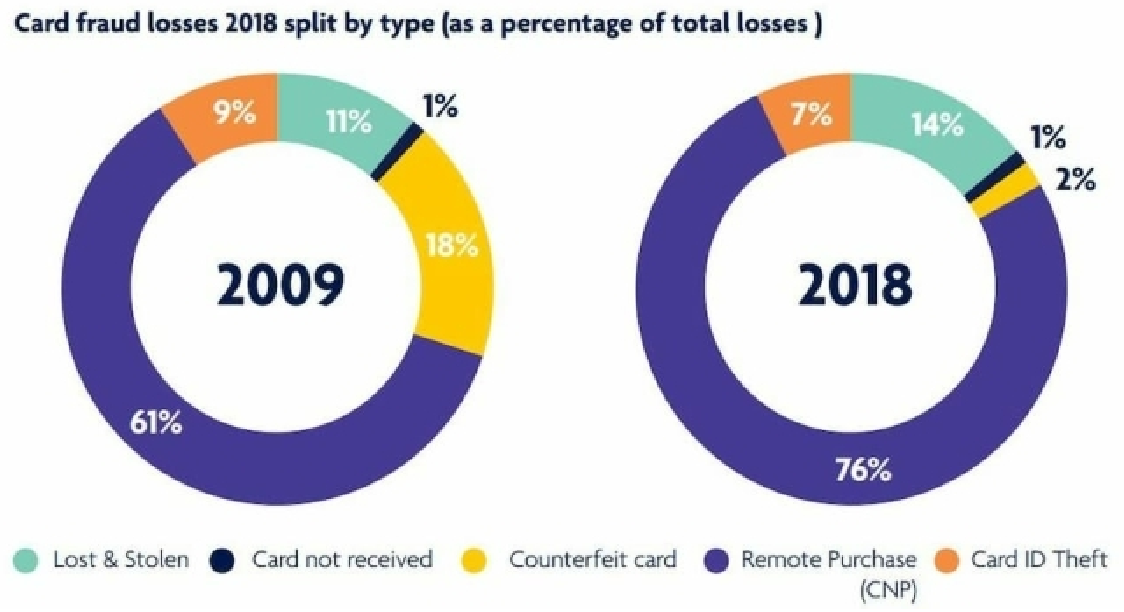
Consumer Analytics

Algorithmic Trading

Fraud Detection in Financial Services

Detecting and fending off credit card fraud become increasingly important

What is Credit Card Fraud Detection?
Credit card fraud detection is the process of identifying purchase attempts that are fraudulent and rejecting them rather than processing the order.



Unauthorized card operations hit an astonishing amount of **16.7 million victims** in 2017. Additionally, as reported by the Federal Trade Commission (FTC), the number of credit card fraud claims in 2017 was **40% higher** than the previous year's number. There were around **13,000** reported cases in California and 8,000 in Florida, which are the largest states per capita for such type of crime. The amount of money at stake will exceed **approximately $30 billion by 2022**.

Problem Statement

- From the moment the e-commerce payment systems came to existence, there have always been people who will find new ways to access someone's finances illegally. This has become a major problem in the modern era, as all transactions can easily be completed online by only entering your credit card information.
- It is important that credit card companies are able to **recognize fraudulent credit card transactions** so that customers are not charged for items that they did not purchase.
- Fraud can be committed in different ways and in many industries. The majority of detection methods combine a variety of fraud detection datasets to form a connected overview of both valid and non-valid payment data to make a decision.

Data and Exploratory Analysis

Data Sources

The dataset contains credit card transactions but is highly unbalanced

Key Attributes

- Transaction date and time
- Transaction location
- Unique transaction id
- Merchant information (name and location)
- Category
- Amount
- Customer information (name, gender, address, DOB, etc.,)
- Fraudulent transaction (Yes/No)

The dataset contains transactions made by credit cards in 2019 and 2020 by US consumers

This dataset presents transactions that occurred in consecutive 24 months, where we have ***9,651*** frauds out of ***1,852,394*** transactions.

The dataset is highly unbalanced, the positive class (frauds) account for only ***0.52%*** of all transactions

Given the class imbalance ratio, we recommend using **SMOTE** analysis to over sample the minority class

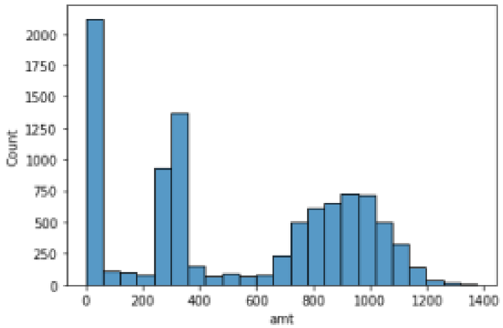
Exploratory Analysis

First glance at the data

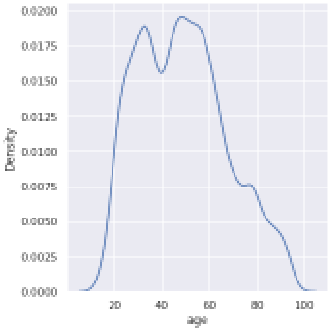
- **~693 merchants**
  - df['merchant'].nunique()
- **~1000 customers**
  - (df['first'] + ' ' + df['last']).nunique()
- **No missing values**
  - df.isna().sum()
- **~1.9 million rows, 22 columns**
  - df.shape()
- **Binary y: 'is_fraud'**
  - df['is_fraud'].value_counts()

Exploratory Analysis

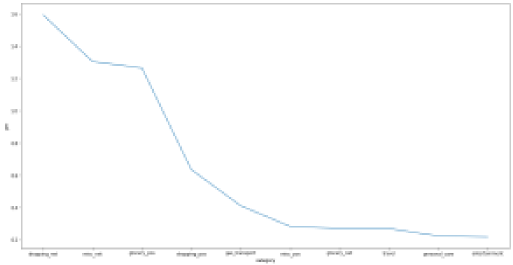
Distributions of key variables are closely examined



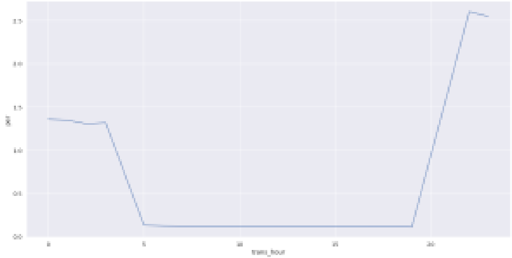
Amount distribution –
Most fraudulent records see transaction amount below $40



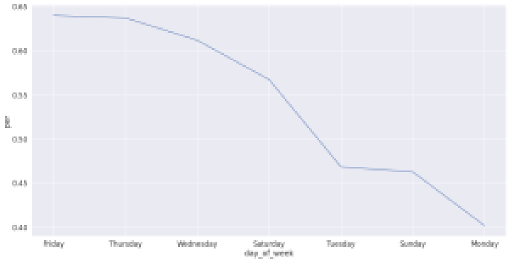
Age distribution –
Higher density of fraud transactions is concentrated with people aged between 21 and 64



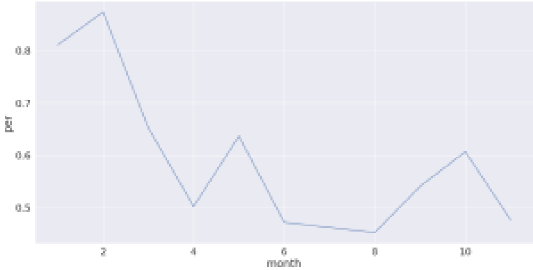
Category distribution –
Internet shopping constitutes most number of fraud transactions



Time distribution –
Sharp increase in number of fraud transactions are recorded after 8pm



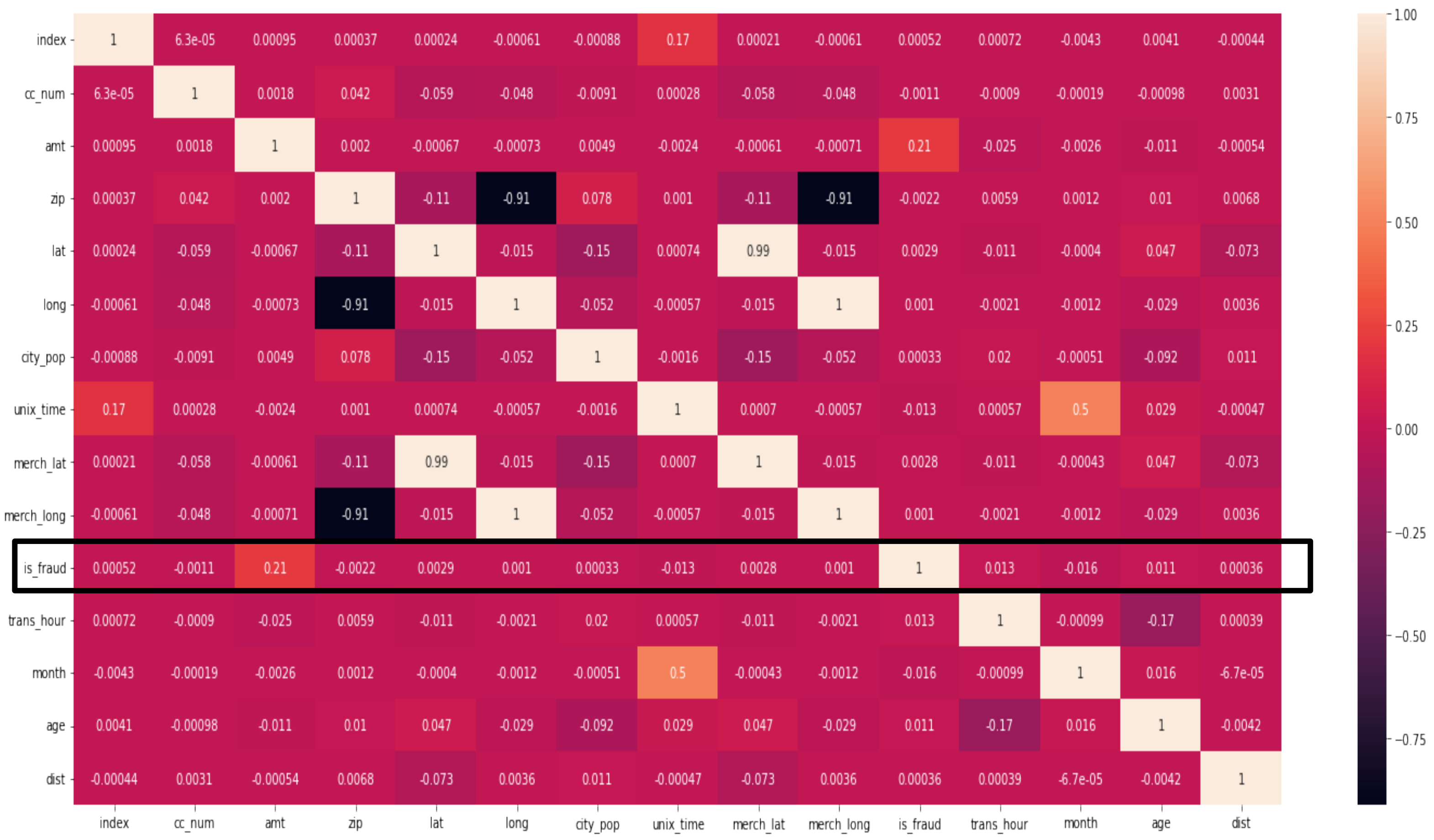
Weekday distribution –
Friday has the highest number of fraud transactions, while Monday has the lowest



Month distribution –
February has the highest number of fraud transactions, while August has the lowest

Exploratory Analysis

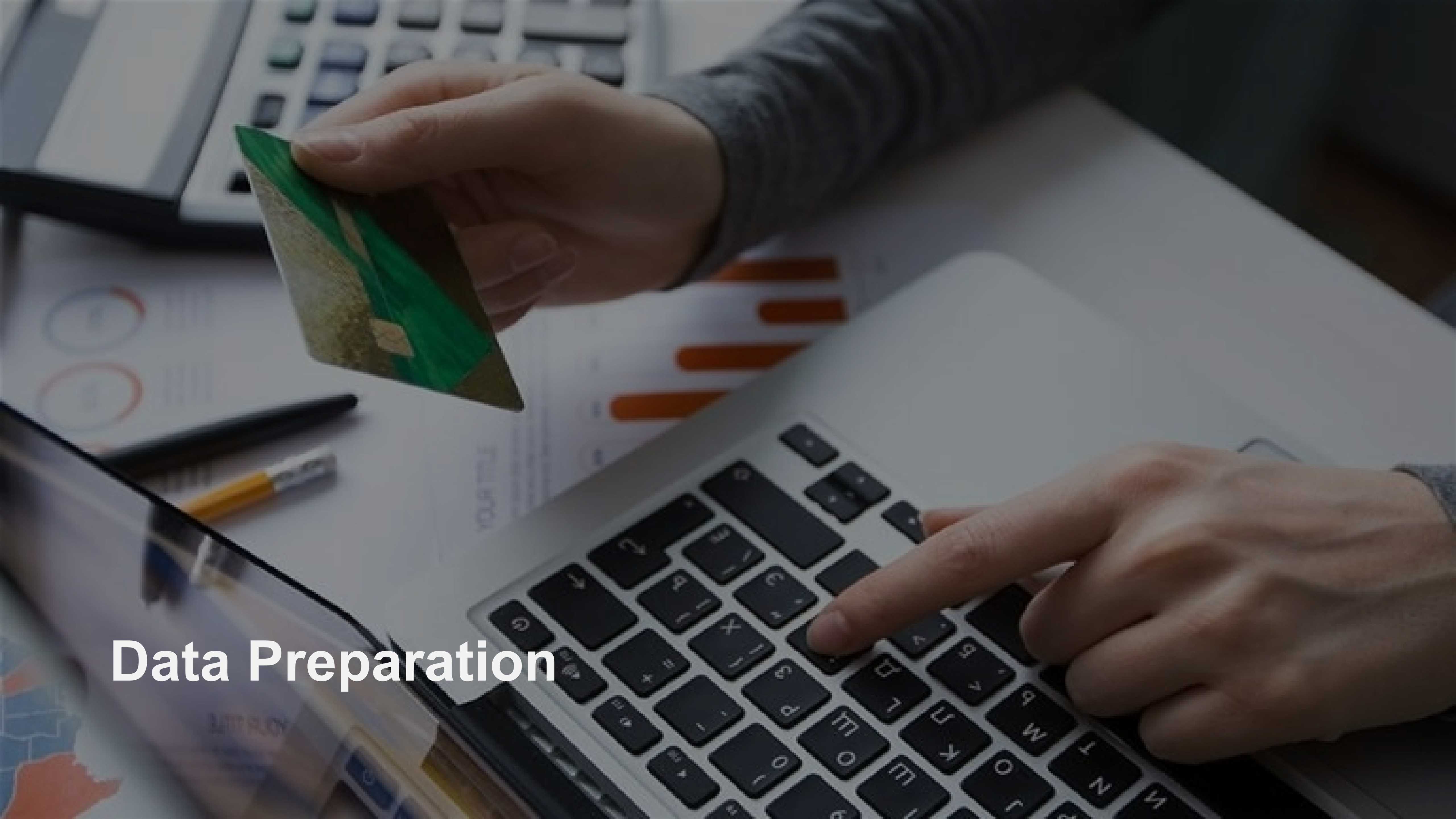
Transactional amount is shown to be the most correlated feature



Correlation Heatmap

Analysis of the correlations indicates:

- The occurrences of fraud transactions are most correlated with transaction amount, time, hour and month
- This indicates that frauds can be explained by the transaction amount and time-related variables in many cases



Data Preparation

Data Processing

One-hot encoding and resampling are performed

01

One-Hot Encoding

Many categorical features are transformed into numerical features to be considered in the ML models (e.g. transaction category, gender, day of the week, etc.)

Resampling

As previously mentioned, the distribution of target variable (is_fraud) is highly unbalanced. After train and test split, resampling is performed on training dataset to make the distribution more even

02

% of fraud transactions in all transactions

0.52% → 20X → 10%

Feature Engineering

To improve the quality of results from a machine learning process

- Transformed raw data into features that can be used in training the model
- Some of the features that were developed include - “60 days transactions count of a customer”, “24 hour transactions count”, “24 hour fraud transactions count”, “2 hour fraud transactions count”, and “60 days average transactions amount”

	is_fraud
is_fraud	1.000000
hist_fraud_trans_24h	0.772578
amt	0.209307
hist_trans_avg_amt_60d	0.084064
hist_trans_60d	-0.047788

It is observed that that ***[Number of fraud transactions in past 24 hours]*** is highly correlated with the target variable. Of the top 4 highly correlated features, three are feature engineered. Indicating the importance of better features to make the model development work well.

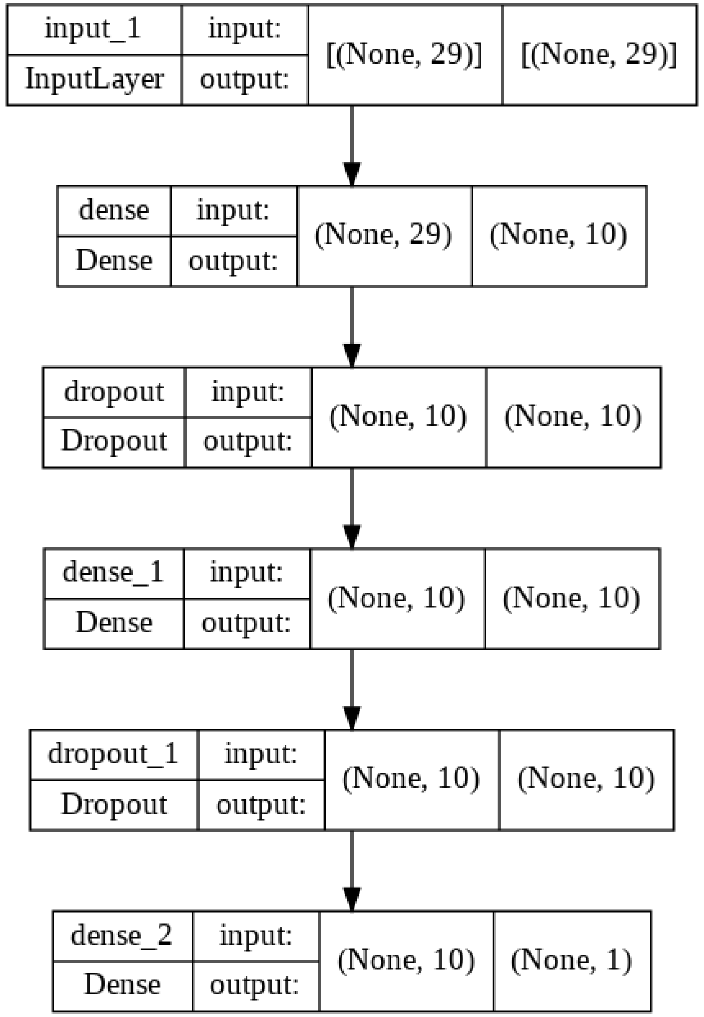
Methodology

Model Selection

Implemented three different models for prediction purposes

Code is executed in Google Colab using a CPU configuration
Grid Search technique is used to find the best model estimators

Logistic Regression	Logistic Regression is used to set up a baseline model while still provide a good performance and model interpretability `(best parameters)  {'C': 1000.0, 'penalty': 'l2'}`
Random Forest	Random Forest as an ensemble method is more robust against overfitting and can potentially produce better results `(best parameters)  {'criterion': 'gini', 'max_depth': 10, 'n_estimators': 10}`
Neural Network	For model architecture used 2 dense layers. Used sigmoid function as activation for the output layer



Findings and Conclusions

Model Results and Evaluation

All models have high accuracy, while Random Forest has high recall rate

Classification Report for Logistic Regression

	precision	recall	f1-score	support
0	1.00	1.00	1.00	552837
1	0.84	0.80	0.82	2908
accuracy			1.00	555745
macro avg	0.92	0.90	0.91	555745
weighted avg	1.00	1.00	1.00	555745

Classification Report for Random Forest

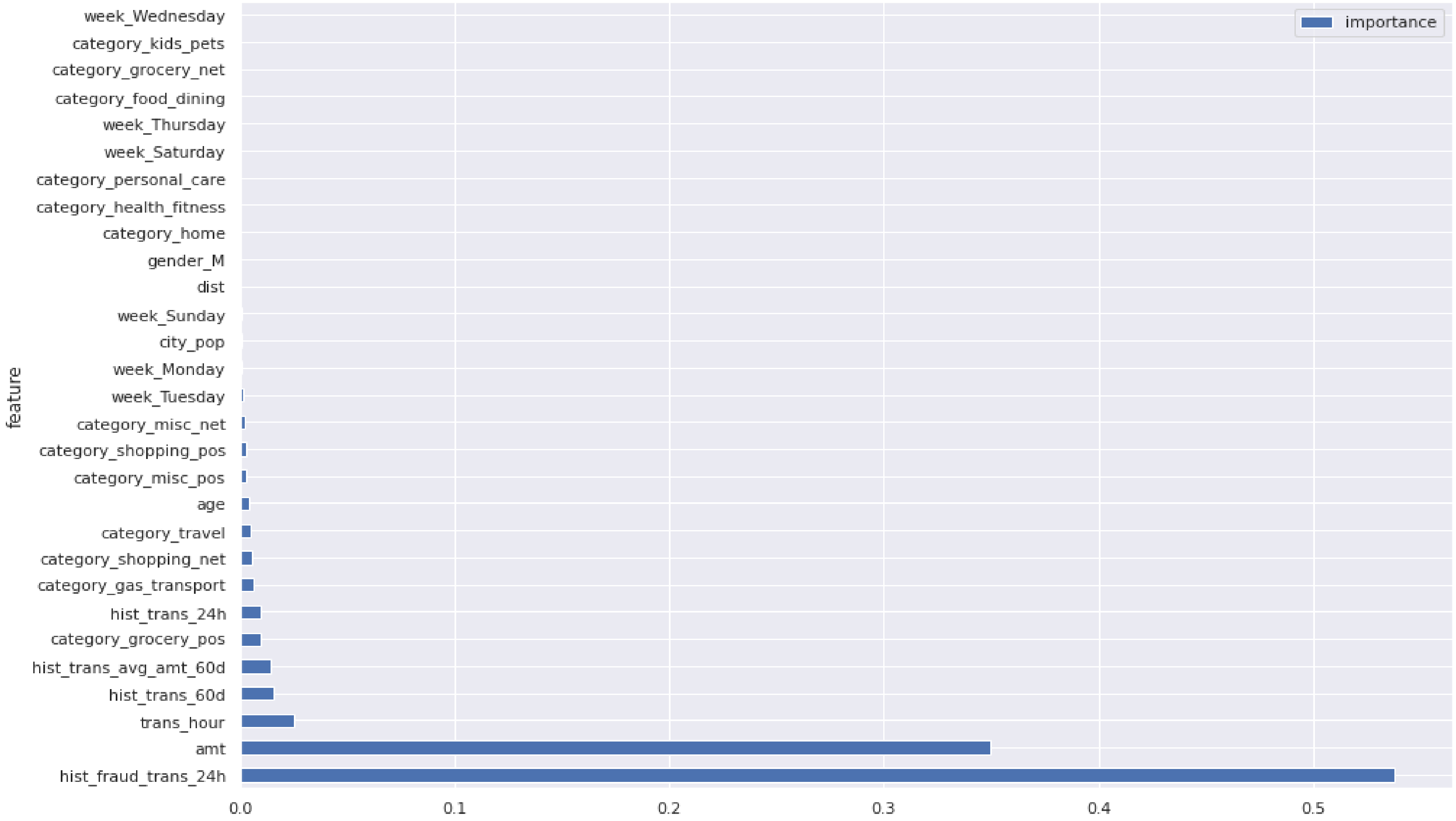
	precision	recall	f1-score	support
0	1.00	1.00	1.00	552837
1	0.97	0.92	0.94	2908
accuracy			1.00	555745
macro avg	0.99	0.96	0.97	555745
weighted avg	1.00	1.00	1.00	555745

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	1.00	0.84	0.80	0.82
Random Forest	1.00	0.97	0.92	0.94
Neural Network	Accuracy: 0.999			

Business Implications

Recent frauds and transaction amounts play a major role in the detection

Feature Importance



According to the feature importance chart:

- ***fraud transaction in past 24 hours*** has the highest importance
- ***transaction amount*** also shares a great deal of importance

In terms of business implications, fraud detection model should focus on a transaction when a fraud transaction has just taken place, or when the amount is relatively high for the user.

Final Thoughts

Lessons Learned and Recommendations

Additional data sources can increase model interpretability given high accuracy

- ➔ We believe our methodology includes a fair combination of models, with baseline logistic regression model, powerful ensemble random forest classifier, and complicated neural networks
- ➔ There can be additional data sources, for example: US holiday calendar can be considered as another layer to the time variables; extra information on victims’ credit card provider can also be useful

Final Recommendation

Although the models have high accuracy, their recall rate and interpretability still have room for improvement. Therefore, more data can be collected to create relevant features and help key stakeholders understand the underlying meanings.

Thank You!

References

https://businessoverbroadway.com/2016/05/09/industry-differences-in-data-science-roles-skills-and-project-outcomes/

https://spd.group/machine-learning/credit-card-fraud-detection/