SafeSwipe
CC Fraud Detection

Brian Woo

Agenda

01 Business Problem

02 Data Analysis

03 Modeling

04 Conclusion

05 Next Steps

06 Demo

01 Business Problem

Financial Institutions are under pressure to combat rising fraud
Fraud harms both finances and customer trust

Issues:

- 60% of U.S. card holders have been victimized by fraud
  - 45% have experienced fraud multiple times

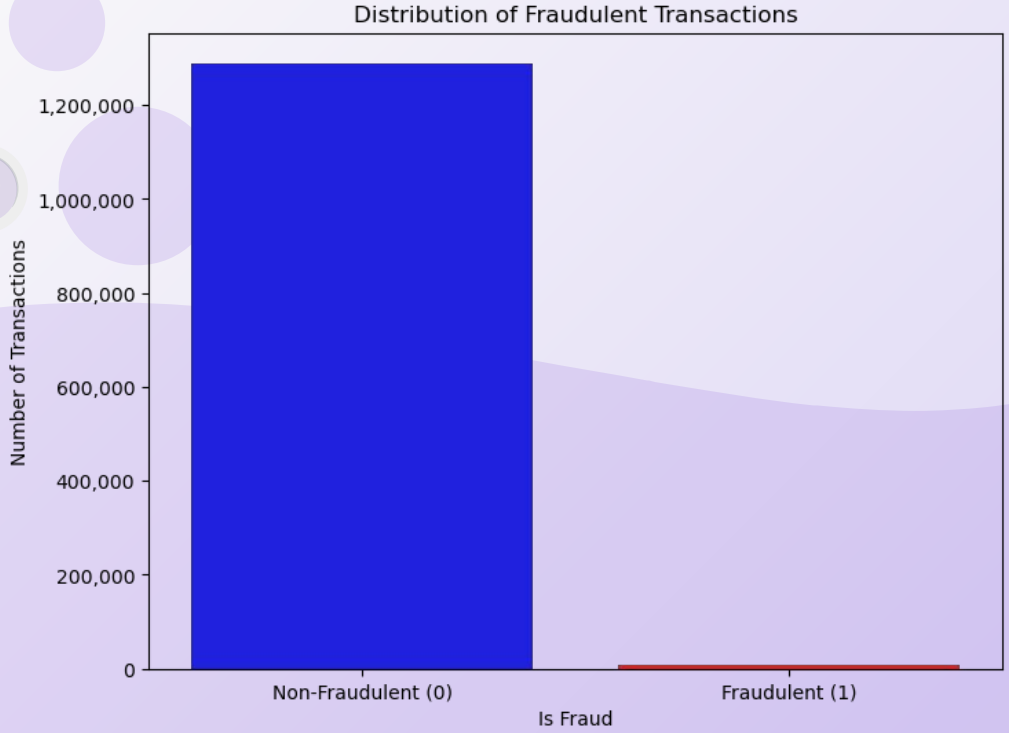
- Last year: 52 million Americans faced fraudulent charges
  - Losses exceed $5 billion

02 Data Analysis

Simulated Dataset contains ~1.3 Million rows of data

- trans_date_trans_time
- category
- amt
- state
- zip
- city_pop
- lat
- long
- merch_lat
- merch_long
- job
- dob
- Target: is_fraud

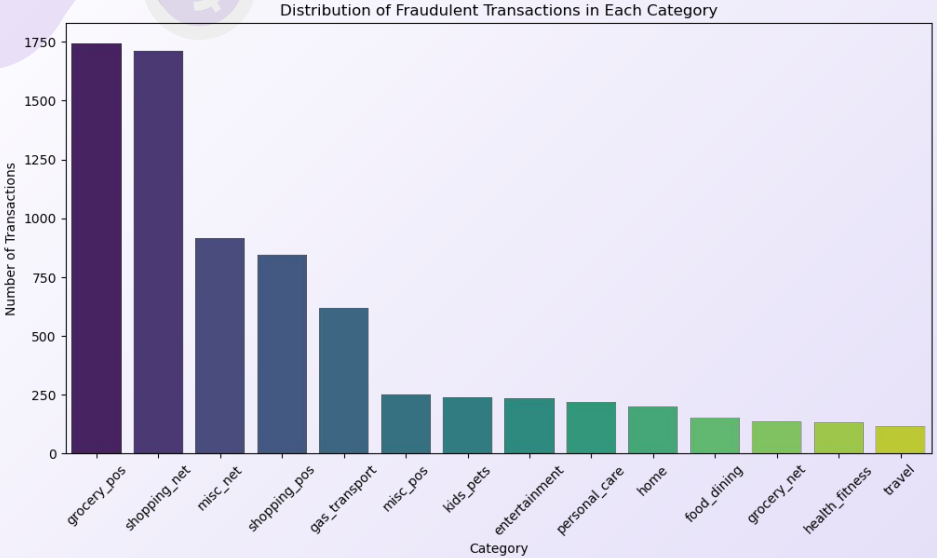
Source: Credit Card Fraud Detection

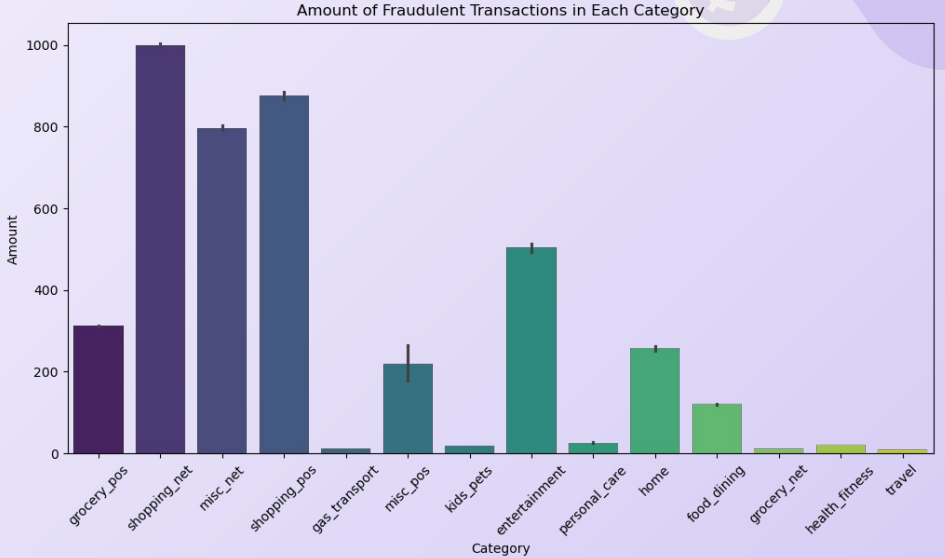


Large Class-Imbalance

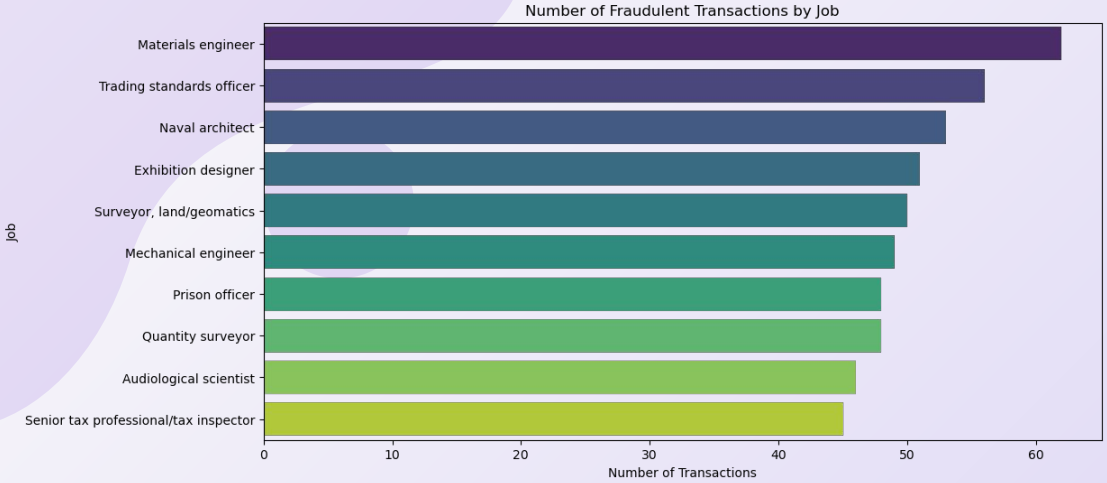
7506 Fraud

Randomly Sampled: 7506 non-fraud

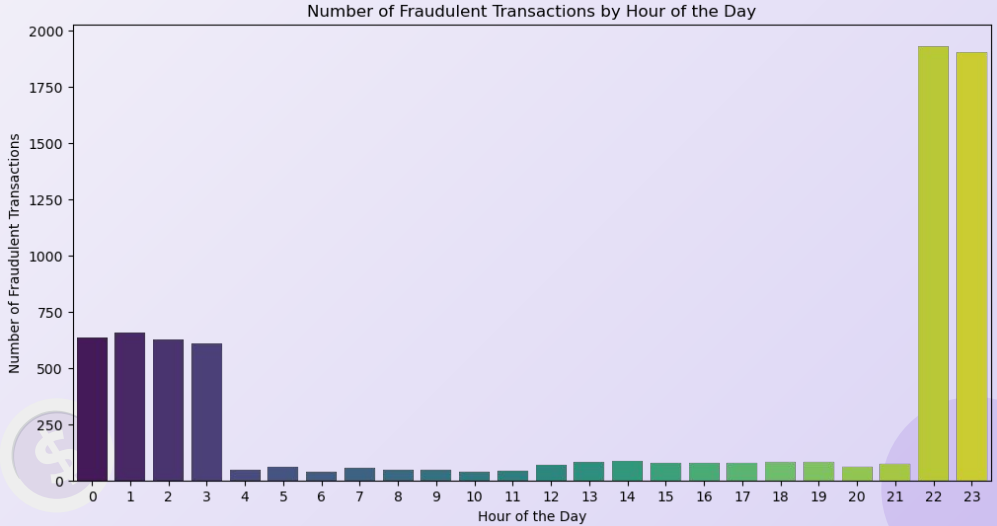




Categories have different spending amounts



Specific jobs have more fraudulent activity



Specific times during the day could indicate patterns for fraud

03 Modeling

Goal: Predict whether a credit transaction is fraudulent or not

Metrics:

- Precision
  - Correctly predicts fraud out of all predicted fraud
  - Reduces false alarms
- Recall
  - Identifies actual fraud cases
  - Catches most fraudulent transactions
- F1-Score
  - Balance between Precision and Recall

Dataset for Training Models

- category
- amt
- state zip
- is_fraud

(Feature Engineered)

- trans_day
- trans_month
- trans_hour
- age_at_transaction
- city_pop_group
- job_category
- time_since_last_trans
- avg_transaction_amount
- category_transaction_count
- unique_transactions_day
- user_merchant_distance

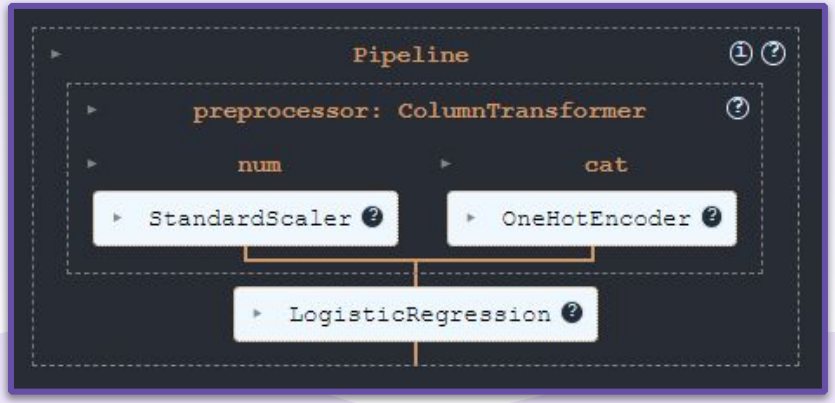
Models:

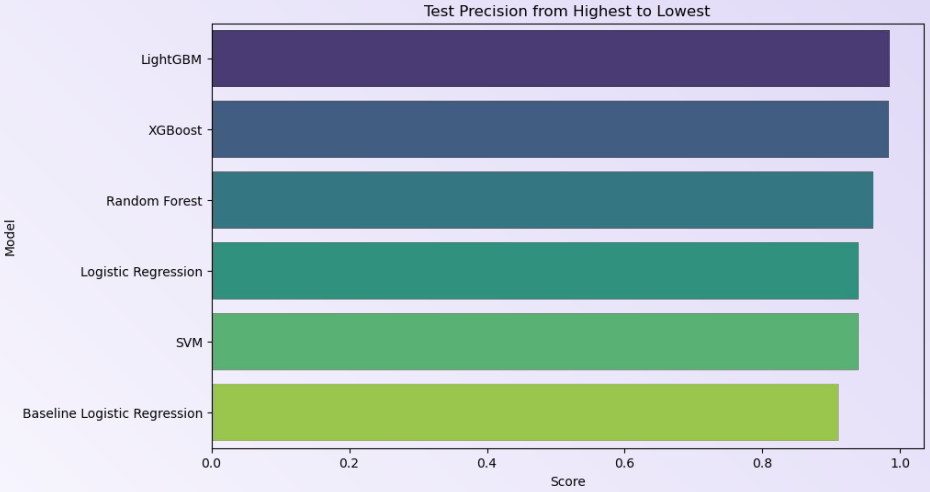
Baseline:

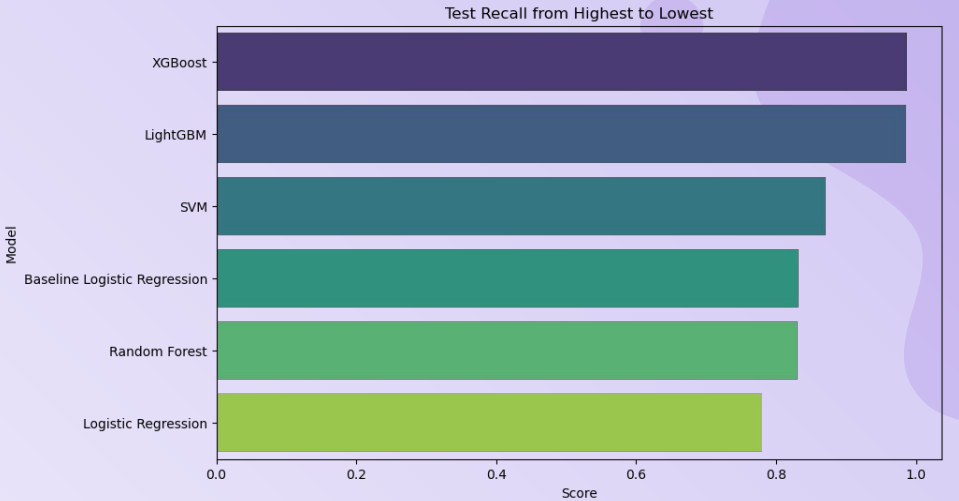
1. Logistic Regression

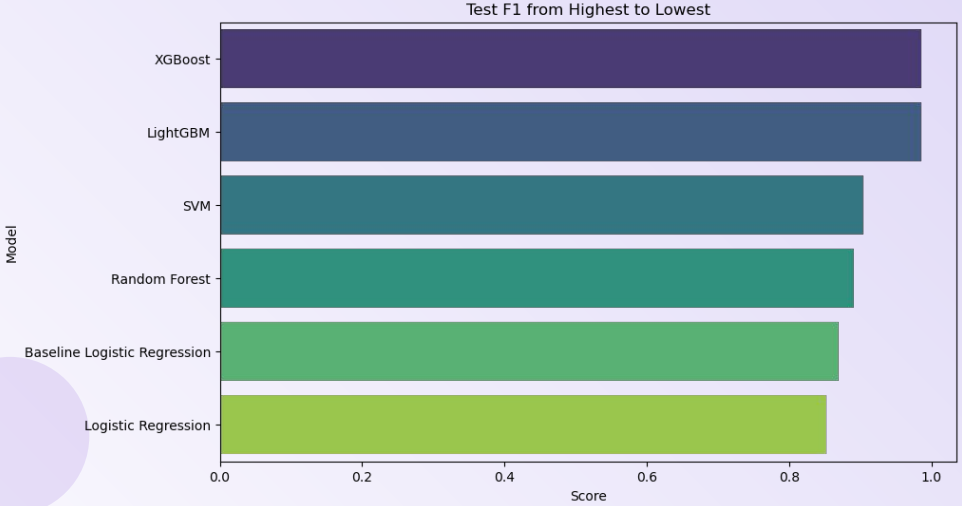
Tuned:

1. Logistic Regression
2. LightGBM
3. XGBoost
4. Random Forest
5. Support Vector Machines




Test Precision from Highest to Lowest
LightGBM
XGBoost
Random Forest
Logistic Regression
SVM
Baseline Logistic Regression
Model
0.0
0.2
0.4
0.6
0.8
1.0
Score


Test Recall from Highest to Lowest
XGBoost
LightGBM
SVM
Baseline Logistic Regression
Random Forest
Logistic Regression
Model
0.0
0.2
0.4
0.6
0.8
1.0
Score


Test F1 from Highest to Lowest
XGBoost
LightGBM
SVM
Random Forest
Baseline Logistic Regression
Logistic Regression
Model
0.0
0.2
0.4
0.6
0.8
1.0
Score

04 Conclusion

	Model	Train Precision	Test Precision	Train Recall	Test Recall	Train F1	Test F1
3	LightGBM	0.998475	0.984472	0.997145	0.985346	0.997810	0.984909
2	XGBoost	0.993143	0.983621	0.992387	0.986679	0.992765	0.985147
4	Random Forest	0.967514	0.960493	0.833270	0.831261	0.895388	0.891216
1	Logistic Regression	0.942632	0.940011	0.775600	0.779307	0.850997	0.852149
5	SVM	0.932290	0.939597	0.872668	0.870337	0.901494	0.903642
0	Baseline Logistic Regression	0.907161	0.910593	0.829463	0.832149	0.866574	0.869606

Best Model: XGBoost

- 2nd highest Precision score
  BUT
- Highest Recall and F1-Scores
- Highest Overall Scores

05 Next Steps

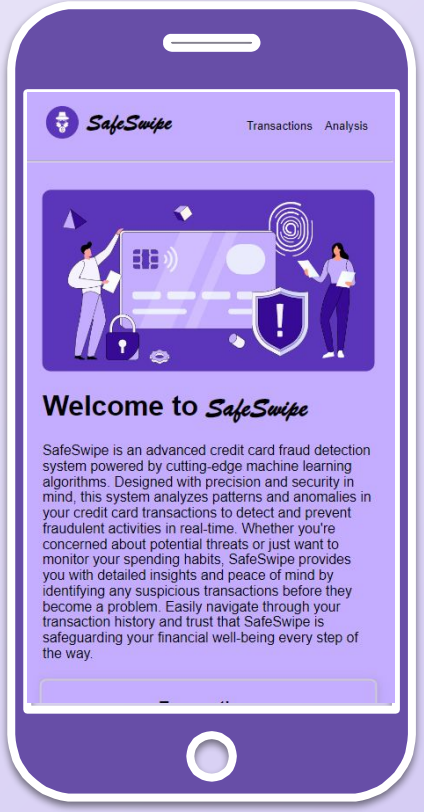
Financial Institutions

- Build Account Database: managing user accounts
- Flexible Input: allow optional fields for easier use
- Implement Real-Time Alerts

06 Demo

Check out our app!

https://safeswipe-e7d39aac3b48.herokuapp.com/



THANKS!

Do you have any questions?
brianhwwoo@gmail.com
https://github.com/Haoweee