Developing a Fraud Detection Model for Risk Management

March 24, 2024

Created by: Ya Min Thu

Agenda

- Introduction to Fraud Detection
- Understanding Fraud Dynamics
- Challenges in Fraud Detection
- The Role of Data Science in Fraud Detection
- Data Collection and Preparation
- Developing the Fraud Detection Model
- Model Implementation Plan
- Performance Metrics for Fraud Detection
- Manual Investigations and Ground Truth Establishment
- Model Monitoring and Maintenance
- Handling False Positives and Negatives
- Case Studies: Successful Fraud Detection
- Conclusion and Next Steps
- References

Introduction to Fraud Detection

01

Fraud detection is crucial for risk management to identify and prevent potentially fraudulent activities.

02

Early fraud detection can mitigate financial losses and protect the organization's reputation.

03

The goals of the presentation include defining fraud detection, emphasizing early intervention, and setting the stage for developing a fraud detection model.

Understanding Fraud Dynamics

Types of Fraud

- Identity theft
- Credit card fraud
- Account takeover

Indicative Patterns

- Unusual transaction times
- Sudden high-value transactions
- Multiple failed login attempts

Challenges in Fraud Detection

1

Developing a fraud detection model to identify complex fraud patterns effectively.

2

Limited resources for manual investigations, restricting ground truth establishment.

Class imbalance in dataset.

3

Balancing accuracy and speed in fraud detection processes.

4

Need for efficient solutions to optimize fraud detection with limited resources.

Performance need to be optimized for the imbalance production data.

The Role of Data Science in Fraud Detection

Identifying Fraud Patterns

Data science enables the detection of complex fraud patterns through advanced analytics and algorithms.

Predictive Modeling and Machine Learning

Utilizing predictive modeling and machine learning algorithms enhances the accuracy and efficiency of fraud detection systems.

Benefits of Data-Driven Decisions

Data-driven decisions in fraud detection lead to proactive risk management, cost savings, and improved operational efficiency.

Data Collection and Preparation

Sources of Transactional Data

Kaggle Dataset: Fictional data simulated to mimic real life scenarios

Data Cleaning and Preparation

Detail the steps involved in cleaning data, handling missing values, outliers, and formatting for analysis and modeling.

Importance of Quality Data

Important to have a complete, and reliable data for building a robust fraud detection model.

Balance the dataset by downsample the non-fraudulent transitions as needed to hit our target rate

Data Quality Assurance

Important to have unbiased model evaluation, in order to have good performance in both data validation and production verification processes.

Developing the Fraud Detection Model

Model Development Process

Key steps

Data preprocessing: numerical and categorical data

Feature selection: Relevant features to determine the fraud

Model evaluation and fine-tuning: Search best params for model to get best accuracy

```
data_train.head()
```
✓ 0.0s

	category	cc_num	amt	city_pop	lat	long	job	is_fraud	age_range
0	misc_net	2703186189652095	4.97	3495	36.0788	-81.1781	Psychologist, counselling	0	11_20
1	grocery_pos	630423337322	107.23	149	48.8878	-118.2105	Special educational needs teacher	0	11_20
2	entertainment	38859492057661	220.11	4154	42.1808	-112.2620	Nature conservation officer	0	11_20
3	gas_transport	3534093764340240	45.00	1939	46.2306	-112.1138	Patent attorney	0	11_20
4	misc_pos	375534208663984	41.96	99	38.4207	-79.4629	Dance movement psychotherapist	0	11_20

```
data_train.shape
```
✓ 0.0s

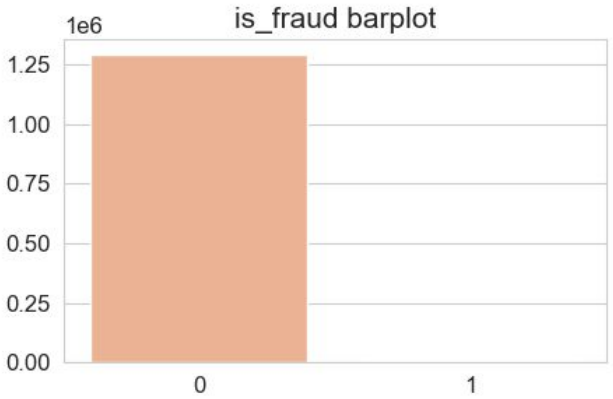
```
(1296675, 9)
```

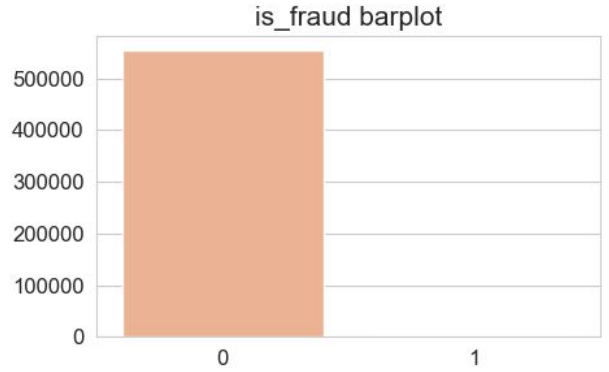
```
data_test.head()
```
✓ 0.0s

	category	cc_num	amt	city_pop	lat	long	job	s_fraud	age_range
0	personal_care	2291163933867244	2.86	333497	33.9659	-80.9355	Mechanical engineer	0	11_20
1	personal_care	3573030041201292	29.84	302	40.3207	-110.4360	Sales professional, IT	0	11_20
2	health_fitness	3598215285024754	41.28	34496	40.6729	-73.5365	Librarian, public	0	11_20
3	misc_pos	3591919803438423	60.05	54767	28.5697	-80.8191	Set designer	0	11_20
4	travel	3526826139003047	3.19	1126	44.2529	-85.0170	Furniture designer	0	11_20

Model Development Process

Imbalance Dataset





Training data: 1296675 Row

Test Dataset: 555719 Rows

Algorithm Selection: Why XGBoost for Fraud Detection

Advantages for Fraud Detection

- Efficient handling of **large volumes** of transaction data.
- Effective management of **imbalanced datasets** with few fraudulent cases.
- **Prevention of overfitting** to generalize well to unseen fraudulent patterns.
- Ability to identify important features contributing to fraudulent behavior.
- Interpretability for explaining and understanding model decisions.

Comparison with Other Models

Logistic Regression:

Interpretable but may struggle with complex relationships compared to XGBoost.

Random Forests:

Robust, but XGBoost often achieves better performance, especially on imbalanced datasets.

Support Vector Machines (SVM):

Powerful but may require more tuning and be less interpretable compared to XGBoost.

Optimization of Model Performance

Things we can do

Feature Engineering:

Enhance model discriminative power by crafting or transforming features through techniques such as binning, scaling, or creating interaction terms.

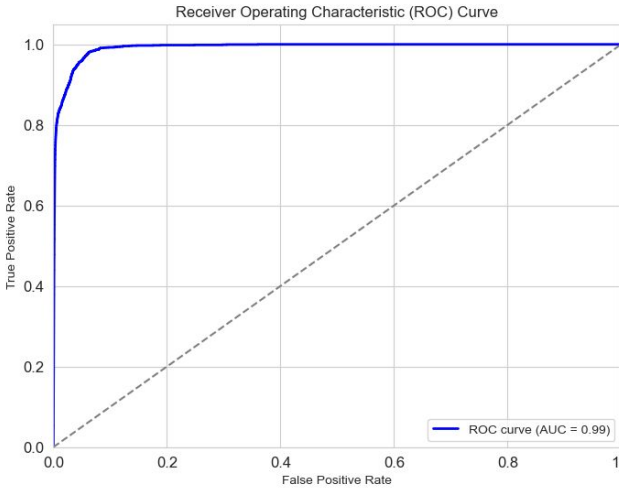
Address Class Imbalance:

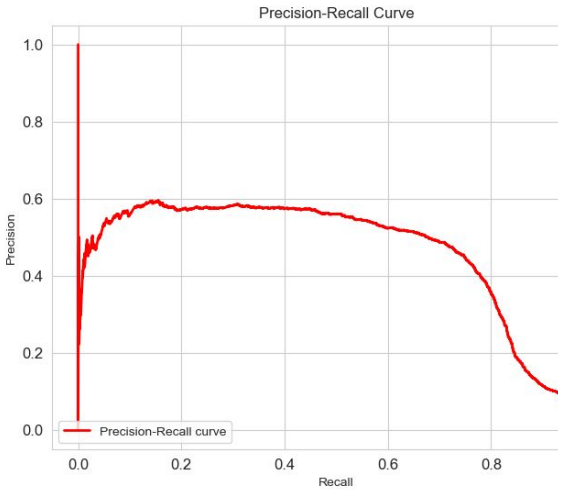
Mitigate class imbalance by employing strategies like oversampling the minority class or leveraging algorithms designed to handle imbalanced datasets, such as XGBoost's scale_pos_weight parameter.

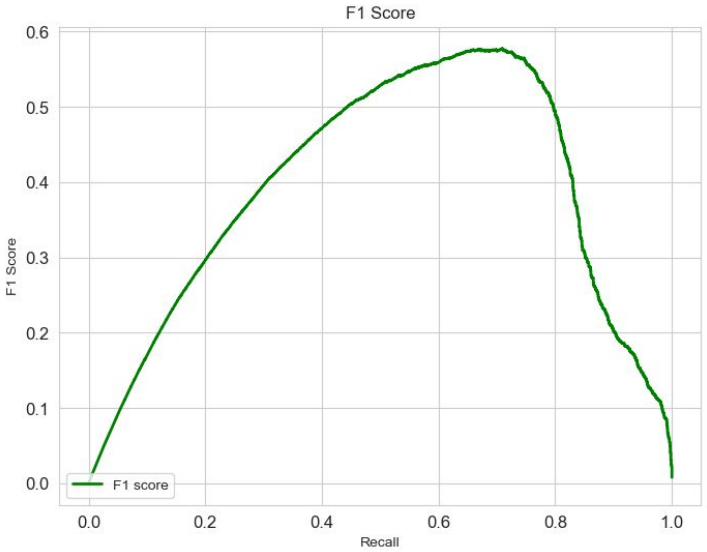
Hyperparameter Tuning:

Fine-tune model performance by exploring hyperparameter spaces more comprehensively, utilizing methods like random search or Bayesian optimization in addition to grid search.

Optimization of Model Performance







Performance Metrics for Fraud Detection

	Training Set	Validation Set	Test Set	Best Model After on Test Set
Accuracy	97.9%	97.2%	95.7%	96.4%
Precision	97.4%	96.7%	7.9%	9.19%
Recall	98.4%	97.5%	94.9%	92.2%
F1 Score	97.9%	97.2%	14.6%	16.7%

Low recall and F1 score indicate that the model is failing to effectively identify a significant portion of the positive instances (e.g., fraudulent transactions) while also struggling to achieve a balance between precision and recall.

Model Implementation Plan

Deployment Steps

Prepare the model for production deployment.

Test the model in a controlled environment for validation.

Deploy the model to the operational system for real-time fraud detection.

Integration with Transaction Systems

Integrate the fraud detection model with existing transaction processing systems.

Ensure seamless data flow and compatibility between the model and operational systems.

Monitor system performance post-implementation for optimization

Manual Investigations and Ground Truth Establishment

Manual investigations are crucial for training the fraud detection model by providing labeled data for learning patterns.

Resource constraints limit manual investigations to 1000 per month, impacting the volume of ground truth data available for model refinement.

Ground truth data from manual investigations serves as a benchmark for evaluating the model's accuracy and improving its performance over time.

Iterative model training using ground truth data enhances the model's ability to detect and prevent fraudulent transactions effectively.

Model Monitoring and Maintenance

Ongoing Monitoring

Regularly assess model performance to ensure accuracy and effectiveness in fraud detection.

Performance Tracking

Utilize key metrics like accuracy, precision, recall, and F1 score to measure model performance over time.

Model Maintenance Strategies

Implement scheduled updates, retraining, and recalibration of the model to adapt to evolving fraud patterns and maintain effectiveness.

Handling False Positives and Negatives

Strategies to Minimize Errors

- Implement threshold tuning to balance precision and recall in the model.
- Utilize ensemble methods to reduce false positives and negatives.
- Regularly update the model with new data to improve accuracy.

Impact on Operations

- False positives can lead to unnecessary investigations, straining resources.
- False negatives may result in missed fraudulent transactions, impacting financial security.
- Inaccurate classifications can erode trust with customers and partners.

Case Studies: Successful Fraud Detection

Credit Card Fraud Detection

Given the global prevalence of credit card fraud, the majority of banks have opted for fraud detection services offered by industry providers, encompassing the prevalent types of fraud.

E-commerce Fraud Detection:

Implement strategies to detect and prevent online fraud, including identifying suspicious payments, detecting new account fraud, preventing abuse of trial and loyalty programs, and improving detection of account takeovers.

Conclusion and Next Steps

- Developed a fraud detection model for early intervention in potentially fraudulent transactions.
- Highlighted the significance of data science in risk management through fraud detection.
- Next steps include model implementation, monitoring, and ongoing optimization.

References

- OCBC. (n.d.). Detecting credit card fraud. OCBC Careers Blog. Retrieved from https://www.ocbc.com/group/careers/blog/detecting-credit-card-fraud.html
- Kaggle. (n.d.). Fraud transactions dataset. Retrieved from https://www.kaggle.com/datasets/dermisfit/fraud-transactions-dataset
- Amazon Web Services. (n.d.). Amazon Fraud Detector. Retrieved from https://aws.amazon.com/fraud-detector/