AUTOMATED INVOICE PROCESSING AND FRAUD DETECTION

Elvis O Antwi

PERSONAL PROJECT https://github.com/Elvis-Opoku

Executive Summary

This project analyzed over 1.1 million invoices and flagged nearly 30,000 that showed signs of fraud. Using a mix of simple business rules and machine learning, it focused on patterns like duplicate orders, unusual credit use, and spikes in specific regions and currencies. Most of the issues showed up in USD and EUR transactions, especially from the South Region and countries like Korea and Congo. The final dashboard makes it easy to explore these risks by time, region, or channel, giving finance teams a practical way to spot and respond to suspicious activity faster.

Business Impact

This project wasn't about building something for it. It was about helping finance teams catch real issues in a sea of transactions. Here's how it makes a difference:

Early fraud detection: Nearly 30,000 invoices were flagged before any money went out the door. That gives finance teams a chance to review and take action before it turns into a bigger problem.

Seeing where problems are coming from: The dashboard breaks down fraud patterns by region, currency, and channel. USD and EUR invoices showed the highest risk, with the South Region and a few countries like Korea and Congo standing out.

Saving time on reviews: Instead of sorting through a million records, the model narrows it down to the few that need a closer look. It's a faster way to spot the red flags.

Making the data easy to work with: The dashboard isn't just numbers on a screen. It's something teams can use to filter by date, region, or channel and see what's happening without needing technical support.

Fewer costly mistakes: By catching duplicates, inflated amounts, or credit misuse, the system helps avoid errors that can lead to financial loss.

Project Overview

This project started with a basic but important question: how do you help a finance team catch invoice fraud without burying them in spreadsheets? With over a million records to review, going line by line just isn't practical. Manual reviews are slow, easy to miss things, and usually catch problems only after they've caused damage.

So, the idea was to build something that does heavy lifting, something that catches both the obvious red flags and the subtle patterns a person might overlook. It combined two approaches: clear business rules to catch things like duplicate invoices or strange credit use, and a machine learning model (Isolation Forest) to pick up on less obvious behavior, like unusual timing or vendor activity.

The results are presented in a Power BI dashboard that anyone can explore filter by region, date, currency, or vendor to find what stands out. It's not just about detecting fraud. It's about helping people see where things might go wrong and giving them the tools to act before they do.

Tools & Technology	
Category	**Tool/Platform**
Data Processing	Python (Pandas, NumPy)
ML/Anomaly Detection	Sklearn (Isolation Forest)
Reporting Dashboard	Power BI
File Format	Excel / CSV / .pbix

Data Overview

The dataset combines real-world invoice records from Kaggle with additional synthetic data to create a richer, more varied sample for fraud detection. In total, there are just over 1.1 million invoices and 25 columns, covering fields like order amount, credit release, currency, and vendor information. A unique customer order ID links both datasets. Some fields, like purchase order type and credit status, had missing values and were handled during preprocessing. The final dataset reflects a mix of typical financial transactions and simulated anomalies, giving the model enough variety to learn meaningful patterns and flag unusual activity.

Methodology

Data Cleaning & Preprocessing

The dataset had mixed formats and missing values, so the first step was to clean it up. Columns with x and y suffixes were merged, currency values were converted from text to numbers, and dates were parsed to compute delivery times. Missing numeric values were filled using the mean, and categorical values with the most frequent entry. This helped standardize everything before modeling.

Feature Engineering

New features were created to capture behaviors that might indicate fraud. These included the number of days between order and delivery, whether the order was placed on a weekend, and the ratio of credit released to the order amount. The order amount was also standardized to USD using a fixed exchange rate table. Flags were added to mark outliers, duplicate orders, and orders with zero amount but high credit.

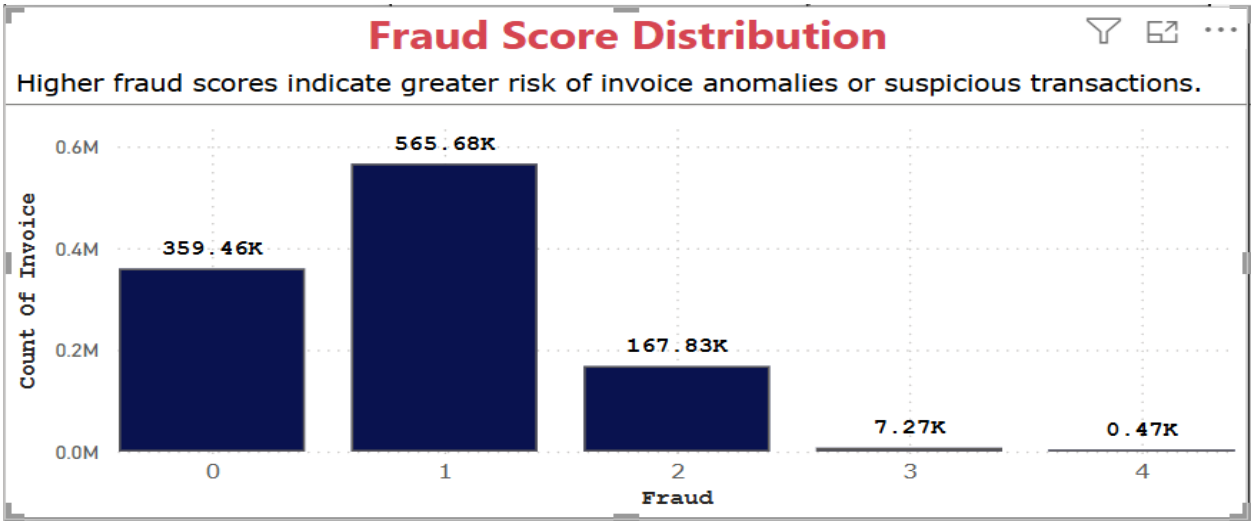
Fraud Detection Models

Two methods were used. The first was rule-based logic: seven simple rules that flagged issues like duplicates or high-risk ratios. Each rule gave a binary score, and the total formed the fraud score. The Isolation Forest machine learning model was subsequently trained on selected numeric features to identify more subtle patterns.

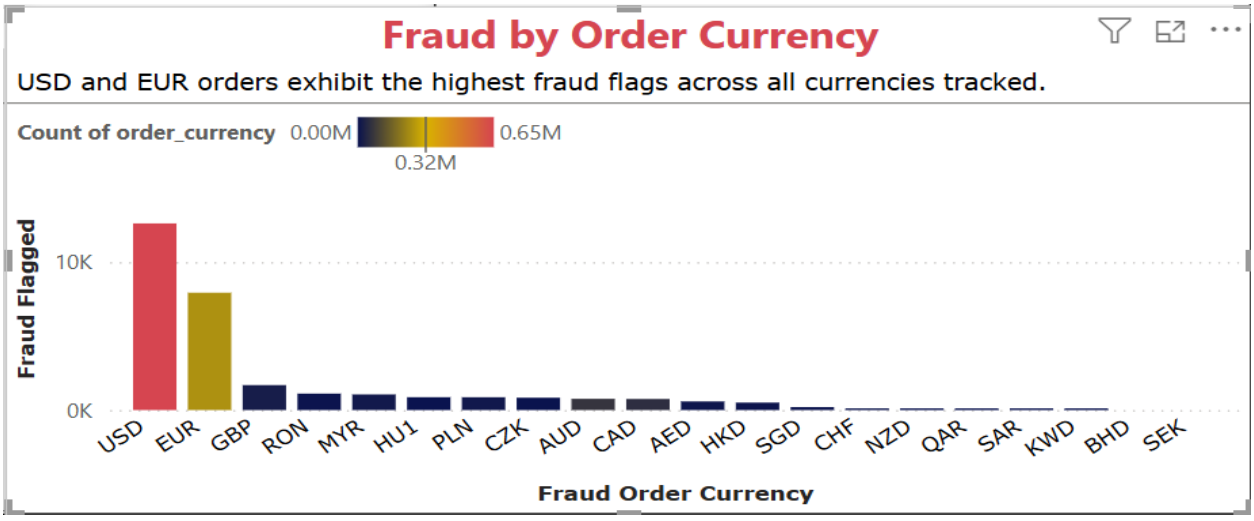
Final Fraud Flag

Any invoice flagged by the model or scoring 3 or more from the rule logic was marked as suspicious. This combined method helped catch both obvious fraud and hidden outliers without overwhelming reviewers with false positives.

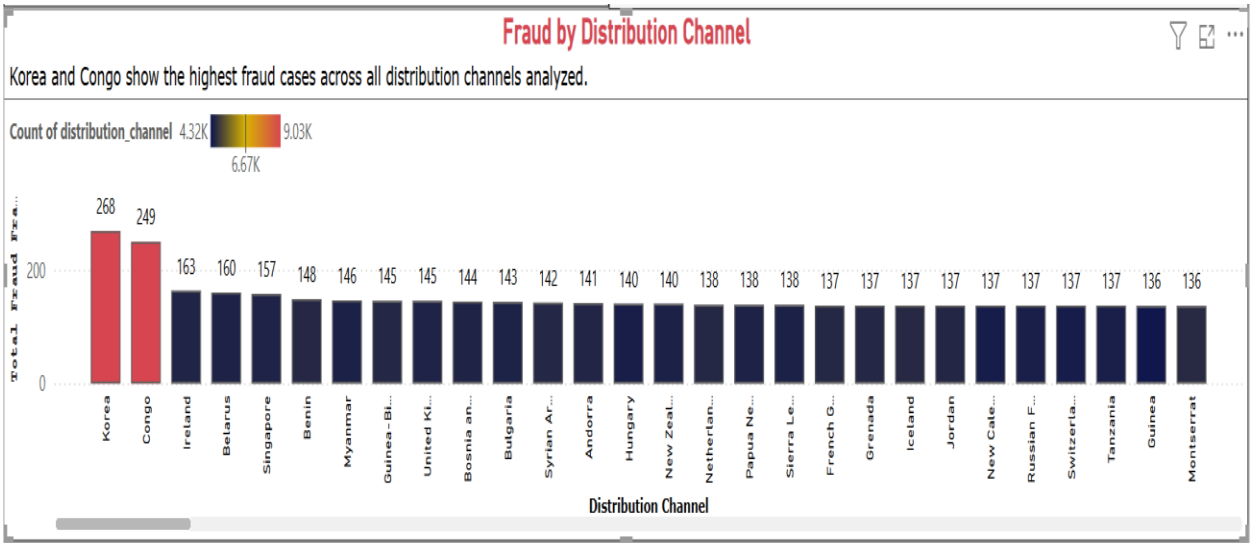
Dashboard Insights



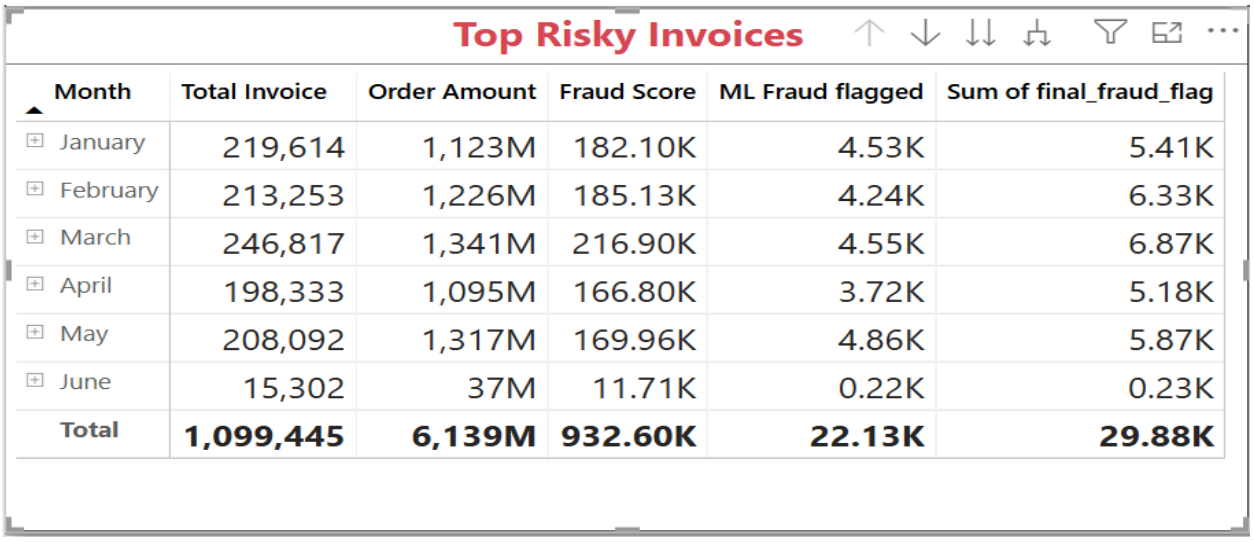
Fraud Score Distribution chart breaks down how many invoices triggered rule-based fraud checks. Most invoices had zero or one issue, but a smaller group had multiple flags, suggesting higher fraud risk. It helps teams quickly see how often suspicious patterns occur and prioritize which invoices need further attention or manual review.



Fraud by Order Currency chart highlights how fraud flags vary by currency. USD and EUR show the highest counts, likely due to higher transaction volumes or inconsistent patterns. This view helps identify which currencies are more vulnerable to suspicious activity and can guide more targeted reviews or currency-specific fraud checks.



Fraud by Distribution Channel chart shows the number of flagged invoices by country or region. Korea and Congo lead, with significantly more fraud cases than other areas. It offers a geographic lens to fraud risk, helping identify regions that may need closer monitoring, better controls, or more frequent review by the audit team.



Top Risky Invoices

Month	Total Invoice	Order Amount	Fraud Score	ML Fraud flagged	Sum of final_fraud_flag
January	219,614	1,123M	182.10K	4.53K	5.41K
February	213,253	1,226M	185.13K	4.24K	6.33K
March	246,817	1,341M	216.90K	4.55K	6.87K
April	198,333	1,095M	166.80K	3.72K	5.18K
May	208,092	1,317M	169.96K	4.86K	5.87K
June	15,302	37M	11.71K	0.22K	0.23K
Total	**1,099,445**	**6,139M**	**932.60K**	**22.13K**	**29.88K**

This table summarizes invoice activity month by month. It includes the number of invoices, total order amount, fraud scores, and flagged cases. March had the highest fraud activity, followed by January and February. The table gives a clear timeline of risk and helps identify monthly patterns worth investigating.

Future Enhancements

Integrate with real-time invoice ingestion APIs: Connecting the system to live invoice feeds would allow fraud detection to happen in real time, rather than after data is uploaded. This means teams can respond faster and potentially stop fraud before any payment is made.

Deploy dashboard on Power BI Service: Hosting the dashboard on Power BI Service would make it accessible to others in the organization without needing to share files. It also supports scheduled refreshes, role-based access, and collaboration through the web.

Implement feedback loop from finance team for ML retraining: Allowing the finance team to identify flagged invoices as true or false positives can enhance the model's accuracy. This feedback could be used to fine-tune the machine learning system over time, making it more accurate.

Conclusion

This project proves that combining simple business logic with machine learning can help finance teams catch invoice fraud more effectively. Instead of relying on slow and often incomplete manual reviews, the solution surfaces high-risk transactions early before money leaves the account. It aims to assist human judgment by highlighting the invoices that require attention. The Power BI dashboard ties everything together in a way that's accessible and easy to use, even for those without a technical background. Going forward, the system can grow integrating live data, incorporating user feedback, and evolving as fraud tactics change. With the

right tools, even massive datasets become manageable, and fraud becomes something teams can stay ahead of not just react to.

References

Almeida, J. C. P., & Romão, M. J. B. (2010). Benefits management for an e-invoice process. *Portuguese Journal of Management Studies*, 15(2), 137–159.

Dragomirescu, O.-A., Crăciun, P.-C., & Bologa, A. R. (2025). Enhancing invoice processing automation through the integration of DevOps methodologies and machine learning. *Systems*, 13(2), 87. https://doi.org/10.3390/systems13020087

Sahu, S., Salwekar, S., Pandit, A., & Patil, M. (2020). Invoice processing using robotic process automation. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6(2), 216–223. https://doi.org/10.32628/CSEIT2062106

Ewen, J. (2020). *Invoice fraud detection using machine learning: An empirical study on anomaly detection with Isolation Forest* [Master's thesis, KTH Royal Institute of Technology]. DiVA. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1461111