Big 4 Financial Risk Insights: An Exploratory Data Analysis

Team members: Jennifer Lopez and Maybel Herrera

IST 718: Big Data Analytics

June 17th, 2025

Project Overview

In this project, we took a closer look at how the Big Four accounting firms (Deloitte, EY, KPMG, and PwC) handled audit risk and compliance between 2020 and 2025. The data gave us insight into important factors such as employee workload, fraud detection, compliance violations, and how each firm was using AI in their auditing process. We wanted to understand not just what the numbers showed, but also what trends were unfolding over time and how new tools like AI were influencing results. This project was about more than just analyzing data. It was about uncovering how firms are responding to risk, technology, and the growing pressure to deliver more accurate and efficient audits.

Prediction Goals:

- Predict the number of high-risk audit cases.
- Estimate potential financial loss per firm.

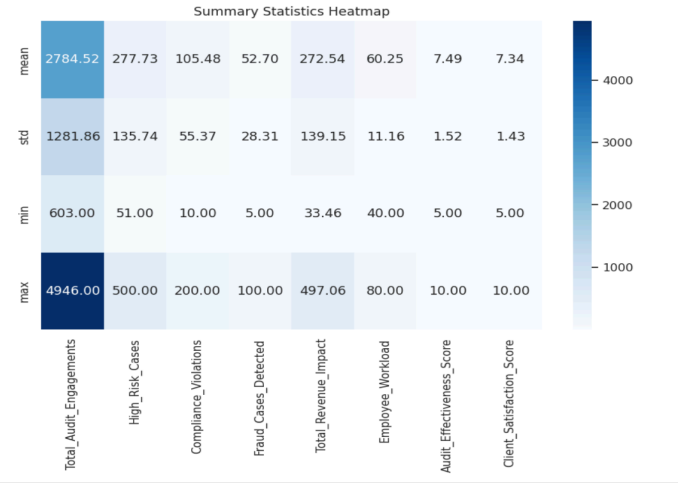
Inference Goals:

- Understand how AI usage affects audit outcomes.
- Explore links between audit volume, workload, and risk.
- Examine how industry differences influence audit-related patterns.

Data Exploration Analysis : The dataset, named big4_financial_risk_compliance.csv, was loaded into a PySpark DataFrame for initial processing and then converted to a Pandas DataFrame for enhanced visualization capabilities.

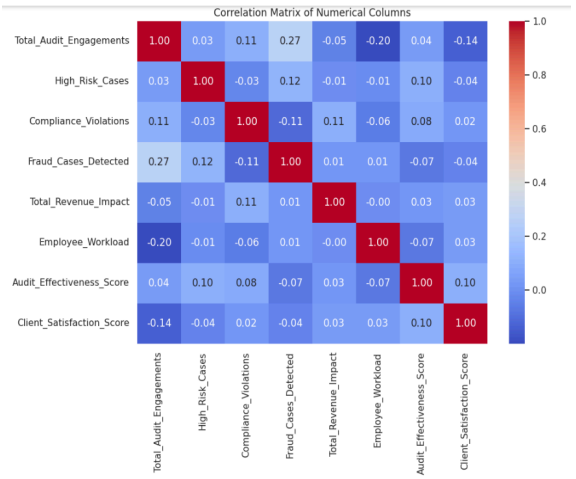
Data Structure and Quality: The dataset contains 100 records and 12 columns. Key variables include firm name, industry, audit volume, high-risk cases, fraud, compliance violations, AI usage, and revenue impact. No missing values were found, ensuring data quality for analysis.

Descriptive Statistics: Summary statistics revealed:



- Audit engagements ranged from 603 to 4,946 (avg. 2,784).
- High-risk cases averaged 278; fraud cases averaged 53.
- Revenue impact ranged from $33M to $497M (avg. $273M).
- Workload ranged from 40–80 hours/week (avg. 60).
- Audit effectiveness and satisfaction averaged around 7.4 out of 10.

Correlation: A correlation heatmap showed weak or moderate relationships among key variables. For example:

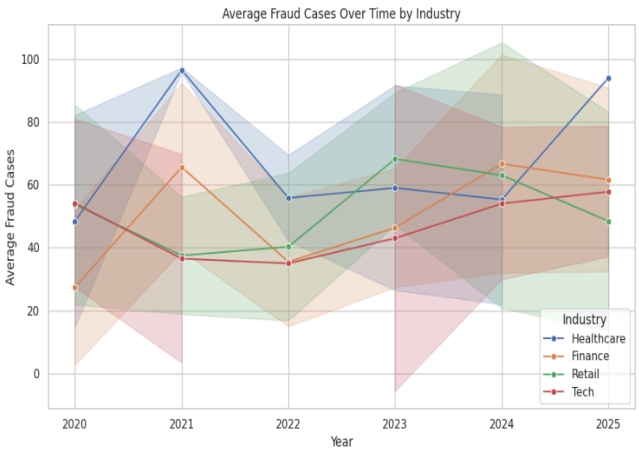


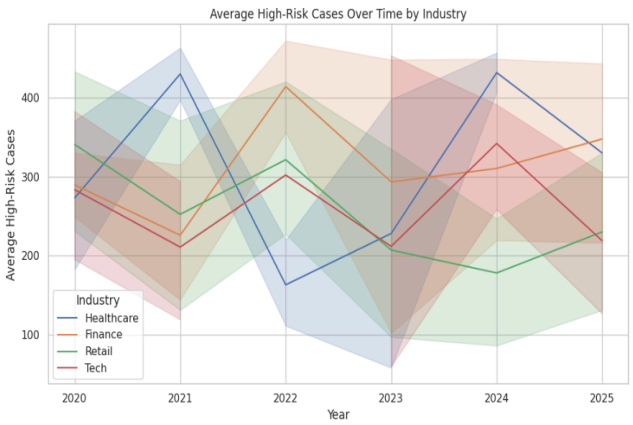
- Fraud cases moderately correlated with audit volume (r = 0.27).
- High-risk cases had weak links to compliance and fraud.
- Audit effectiveness slightly declined with heavier workloads (r = –0.07).
- Client satisfaction improved with audit effectiveness (r = 0.10).

These patterns suggested that linear modeling would not be effective. Audit risk relationships are likely nonlinear and better suited for models like decision trees or random forests.

Trends Over Time: Industry-Level Changes in Fraud and Risk

To explore how audit challenges evolve, average fraud cases and high-risk audit cases were analyzed over time across industries. These trends provide insight into how different sectors experience and respond to audit scrutiny, potential misconduct, and evolving risk exposure.





Fraud Cases: Retail and healthcare showed the most fluctuation in fraud cases, with spikes in 2021 and 2024. Finance remained stable, while tech dipped in 2025. These cycles may reflect policy changes or shifts in audit strategy.

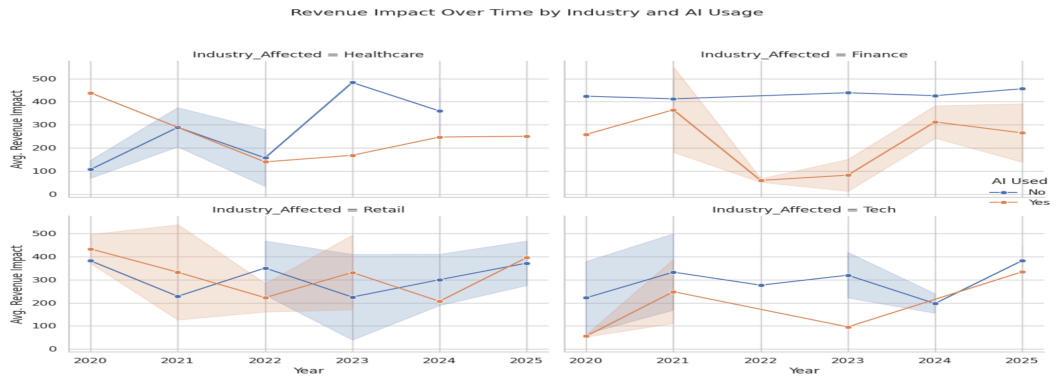
High-Risk Cases: Finance and healthcare consistently had higher average high-risk cases. Tech and retail remained moderate but tech showed a notable jump in 2024.

Exploring the Impact of AI in Auditing

To better understand the role of AI in auditing processes, we examined three key visualizations focused on audit effectiveness, employee workload, and revenue

impact. These visuals reveal patterns over time and across firms and industries, offering insights into how AI adoption may influence outcomes in various dimensions.

Revenue Impact Over Time by Industry and AI Usage



Healthcare:

- Non-AI firms saw large revenue swings, with a major spike in 2023.
- AI-using firms showed a steady upward trend after 2022, suggesting more stability.

Finance:

- Firms without AI had consistently higher revenue impact.
- AI adopters dipped in 2022 but rebounded over time, possibly due to adaptation.

Retail:

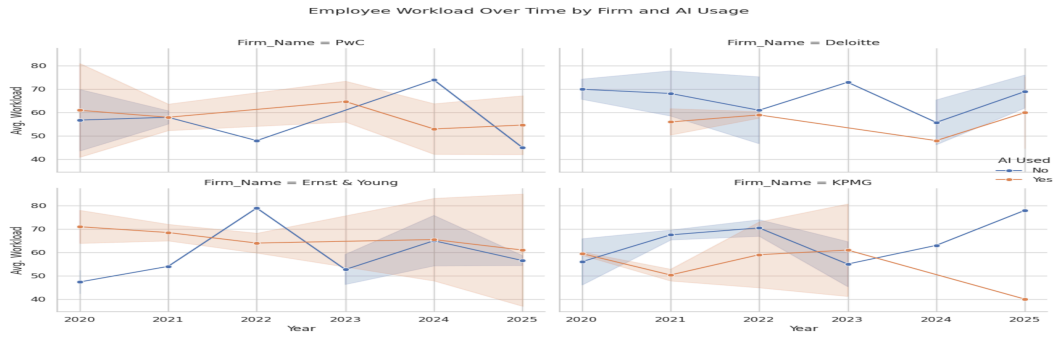
- Showed the most erratic behavior, with AI and non-AI firms repeatedly switching positions.
- Reflects inconsistent AI implementation or external market fluctuations.

Tech:

- AI firms started with lower revenue impact but steadily improved.
- By 2025, they nearly caught up with non-AI firms—highlighting AI's long-term potential.

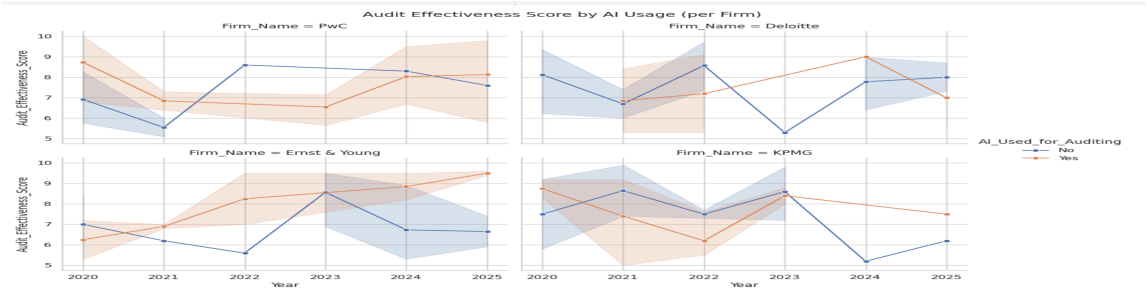
Employee Workload Over Time by Firm and AI Usage

This visualization examines workload trends in the Big Four accounting firms from 2020 to 2025, comparing AI adopters vs. non-AI users.



- **PwC & Deloitte**: AI-equipped teams maintained consistently lower or stable workloads, indicating AI's role in streamlining tasks and reducing manual effort.
- **Ernst & Young:** Non-AI users experienced workload spikes in 2022, while AI users had steadier levels, suggesting AI helps manage peak periods.
- **KPMG**: By 2025, AI-using teams reported the lowest overall workloads, while non-AI counterparts saw a sharp increase highlighting AI's impact on operational efficiency.
- The findings emphasize AI's potential to significantly alleviate workload pressures across major firms

Audit Effectiveness Score by AI Usage (per Firm)



This chart examines changes in audit quality among the Big Four firms, comparing AI adopters vs. non-AI users. PwC & Ernst & Young: AI-supported audits showed a clear upward trend, especially after 2022, leading to more consistent and accurate results.

- **Deloitte**: Some variation occurred, but AI-backed audits still outperformed others, particularly in 2024.
- **KPMG**: Results were the most uneven, yet AI adoption aligned with stronger performance in multiple years, especially early on.
- Findings highlight AI's potential to enhance audit accuracy and consistency across firms

Interesting Results and Patterns

- AI use correlates with improved audit scores and lower workload.
- Retail had fewer high-risk cases but higher financial losses per event
- Finance and healthcare faced frequent risk but with more moderate per-incident costs.
- Tech had fewer incidents overall but still experienced significant financial consequences when failures occurred.

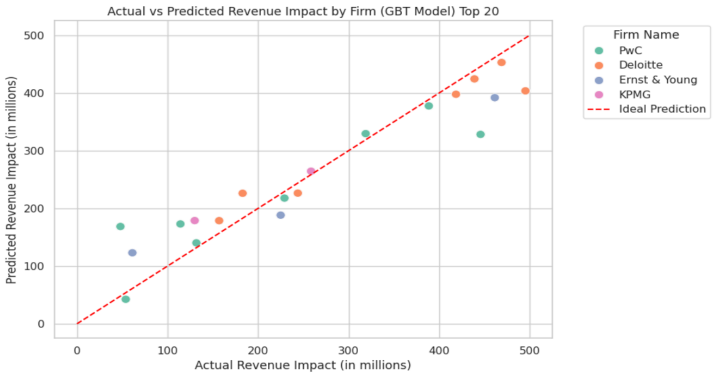
These insights highlight the importance of tailoring risk strategies to each industry's profile.

Predictions and Machine Learning Models :

Can financial losses from audit-related compliance failures be predicted?

Using a Gradient Boosted Tree model, revenue impact was predicted with a Root Mean Squared Error of $66M. Despite variation in actual revenue impact ($33M–$497M), this error margin (~13–20%) reflects strong model performance.

```
GBT Cross-Validated RMSE: 66.13
+------------+---------------+--------------------+------------------+
|   Firm_Name|Industry_Affected|Total_Revenue_Impact|        prediction|
+------------+---------------+--------------------+------------------+
|         PwC|     Healthcare|              114.24|172.98016792335977|
|    Deloitte|     Healthcare|              156.98|178.68737074639904|
|         PwC|     Healthcare|              131.83|140.22939509686805|
|         PwC|     Healthcare|              229.11| 217.80087314245 41|
|         PwC|     Healthcare|               48.0|168.58070038581724|
|    Deloitte|        Finance|              438.89|424.29097098831926|
|    Deloitte|         Retail|              468.82| 452.7583948789777|
|         PwC|           Tech|               53.85|  42.5029733147658|
|         PwC|        Finance|              318.79| 329.4591534710941|
|Ernst & Young|         Retail|              461.33| 391.8003525486932|
|         KPMG|        Finance|              258.49| 264.3882069543531|
|Ernst & Young|     Healthcare|              224.92|188.23394161177353|
|    Deloitte|     Healthcare|              418.49| 397.6633602417752|
|         PwC|         Retail|              445.62| 328.1183472399443|
|         PwC|         Retail|               388.5| 377.4273212849807|
|    Deloitte|        Finance|              495.19|403.57207236185303|
|Ernst & Young|           Tech|               61.17|123.00592758782382|
|    Deloitte|         Retail|               182.9|226.12155189469516|
|    Deloitte|     Healthcare|              243.85|226.36982740983123|
|         KPMG|           Tech|              129.98| 178.8131154964143|
+------------+---------------+--------------------+------------------+
only showing top 20 rows
```

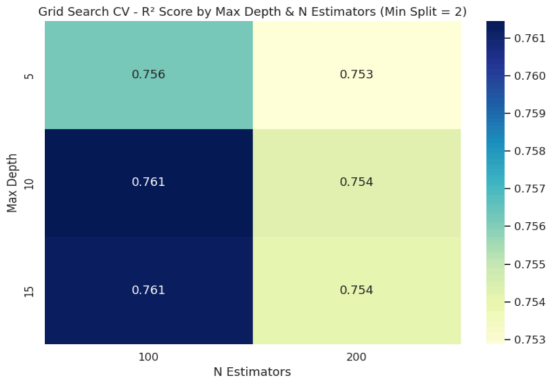


A scatter plot of top 20 cases showed most predictions closely matched actual outcomes. Deloitte and PwC were most often among high-loss cases. KPMG and EY had lower losses and more accurate predictions.

Best Model: Random Forest with Grid Search Optimization

Initial attempts with a Decision Tree gave poor results (R^2 = –0.28). A Random Forest improved the R^2 to 0.20. After applying GridSearchCV and adding engineered features, the optimized Random Forest model achieved:

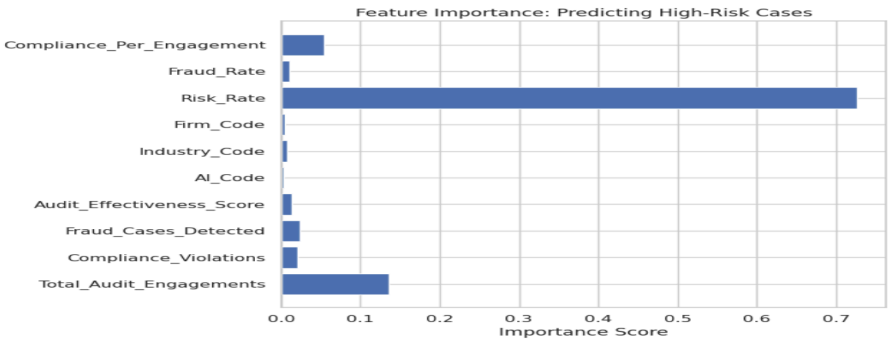
- **R^2 (cross-validation):** 0.76
- **R^2 (test):** 0.712
- **MSE:** 5,104



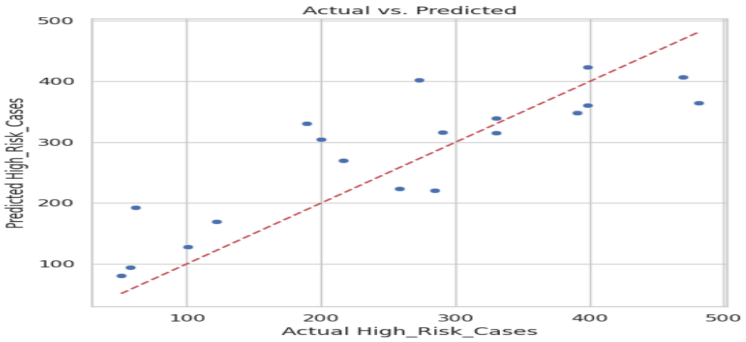
This means the model was able to explain over 71% of the variation in high-risk audit cases, a strong outcome given the complexity of the data.

Feature engineering, especially the creation of Risk Rate, Fraud Rate, and Compliance Per Engagement, was key to this improvement. These variables gave the model deeper context beyond raw numbers.

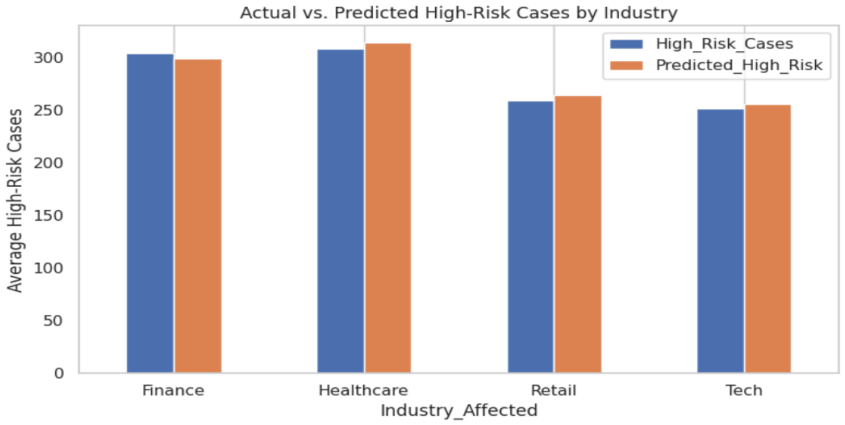
When checking which features mattered most, Risk Rate stood out as the most powerful, followed by Total Audit Engagements and Compliance Per Engagement. Other factors, like which firm or industry didn't play as big of a role.



The scatter plot indicates that the model performs with a fair degree of accuracy, as most data points cluster around the red reference line representing perfect predictions. A clear upward trend emerges, demonstrating that predicted values increase consistently with actual high-risk case counts. While a few observations, particularly in the higher range (300–400), deviate slightly above or below the ideal line suggesting minor over or under predictions.



After validating the model's overall accuracy, the next step was to check whether it remained consistent across different industries. The table below shows the average actual vs. predicted high-risk cases by industry.



Across industries, the model's prediction is closely aligned with actual values. In healthcare, for instance, predicted high-risk cases (314) nearly matched the actual average (308). Retail and tech also showed strong prediction alignment

Problems Encountered : Several challenges emerged:

- Initial correlation analysis revealed weak relationships, making linear models ineffective.
- The first attempt at predicting high-risk cases produced poor results with low predictive power.
- A more advance approach was needed, incorporating tree-based models and targeted feature engineering—was necessary to improve performance.

In conclusion, using data science techniques, this project successfully forecasted high-risk audit cases by experimenting with multiple models and refining features. The Random Forest model emerged as the strongest,

accurately identifying areas of risk and demonstrating the power of machine learning in audit analytics when applied thoughtfully. A key finding was the positive impact of AI adoption, as firms integrating AI into their audit processes achieved more stable outcomes, higher audit effectiveness scores, and reduced employee workloads benefits that enhance both audit quality and team well-being. While fraud prediction remained challenging due to missing behavioral indicators, the study confirmed that meaningful risk patterns could still be captured through the right techniques. Ratio-based features, such as Risk Rate and Compliance Per Engagement, significantly improved model performance by providing essential context beyond raw numbers. These results highlight that predictive analytics, when combined with industry-specific risk patterns, can empower firms to proactively manage compliance and financial exposure. Looking ahead, the ability to tailor strategies based on sector-specific risks will be crucial for organizations aiming to stay ahead in regulatory and financial oversight.

Citations:

Soundankar, Atharva. *Big 4 Financial Risk Insights (2020–2025)*. Kaggle, 2024, https://www.kaggle.com/datasets/atharvasoundankar/big-4-financial-risk-insights-2020-202