Big 4 Financial Risk Insights: An Exploratory Data Analysis

By Maybel Herrera and Jennifer Lopez

Project Overview

Goal:

Explore **audit risk, compliance issues**, and the **impact of AI** across the **Big Four firms (2020–2025)**.

Why It Matters?

Audit failures can lead to **fraud, compliance penalties, and financial losses**.

Firms are adopting **AI-driven auditing**—can it improve accuracy and efficiency

Prediction, Inference and Additional Goals

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.

Prediction Goals:

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

Additional Goals: Used Spark SQL to query the data and answer business oriented questions

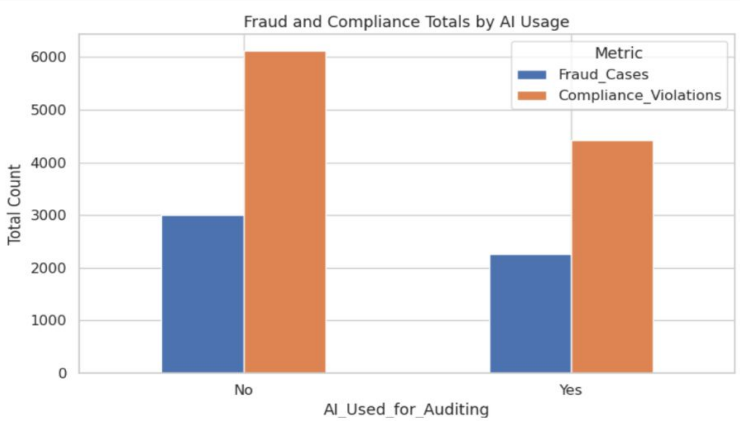
Business Questions ?

Key Questions

- ***Can AI reduce fraud and compliance violations?***
- ***How does AI impact employee workload and audit quality?***
- ***Are certain industries more prone to audit risk?***
- ***Can we predict high-risk audits using historical patterns?***

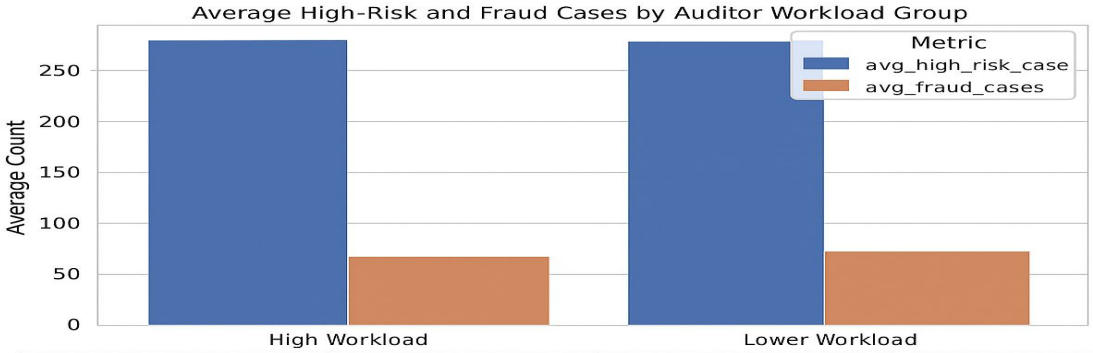
Can AI adoption reduce fraud and improve compliance in Big 4 firms?

- Firms without AI reported 3,010 fraud cases and 6,122 compliance violations.
- Firms with AI reported 2,260 fraud cases and 4,426 compliance violations.
- This reflects a 25% reduction in fraud and a 28% reduction in compliance violations when AI was used.



AI_Used_for_Auditing	total_fraud_cases	total_compliance_violations
No	3010	6122
Yes	2260	4426

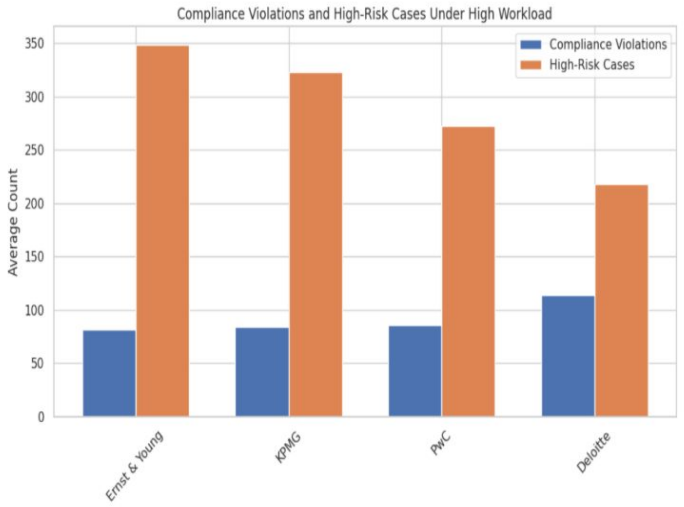
Is there a clear relationship between auditor workload and the likelihood of high-risk audits?



workload_group	avg_workload	avg_high_risk_cases	avg_fraud_cases
High Workload	74.2	282.8	49.7
Lower Workload	54.8	275.8	53.9

- Teams with high workload handled about 283 high-risk cases
- Teams with lower workload handled about 276 high-risk cases

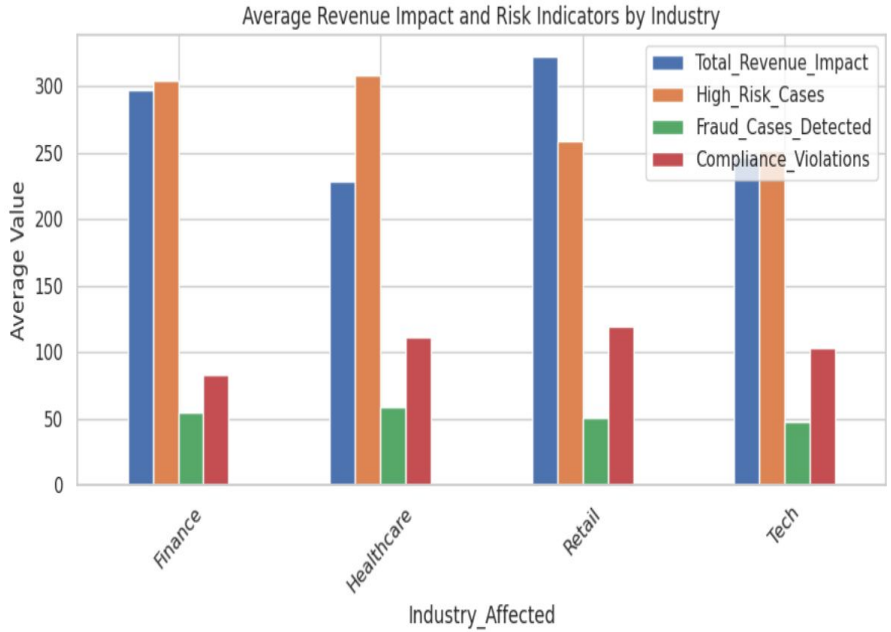
Which firm managed to maintain the best compliance performance despite a high workload?



Firm_Name	avg_workload	avg_compliance_violations	avg_high_risk_cases
Ernst & Young	74.9	81.3	348.3
KPMG	72.8	84.0	323.0
PwC	75.0	85.8	272.6
Deloitte	74.1	114.0	217.9

*Ernst & Young had the highest workload and stands out with the lowest compliance violations** (≈81) among peers under pressure, even though its high-risk case volume is the highest.*

Which industries are most affected by audit-related risk, and how do financial and compliance indicators compare?



- **Retail** faces the highest average revenue impact **(~$322M)**, despite not leading in fraud or compliance violations. This suggests fewer but more financially severe incidents.
- **Healthcare** shows the highest volume of high-risk cases and violations but has a lower average financial impact **(~$228M)**, indicating more frequent, lower-cost issues.
- **Finance** experiences both high-risk exposure and substantial financial losses **(~$297M)**, signaling a dual challenge of volume and cost.
- **Tech** reports fewer fraud and compliance problems but still shows a notable average loss **(~$246M)**, suggesting tech may experience fewer incidents, but the financial impact per failure is substantial.

Data Exploration

About the Dataset

The data set name is big4_financial_risk_compliance.csv.

Data Structure and Quality

- 12 columns
- 100 rows

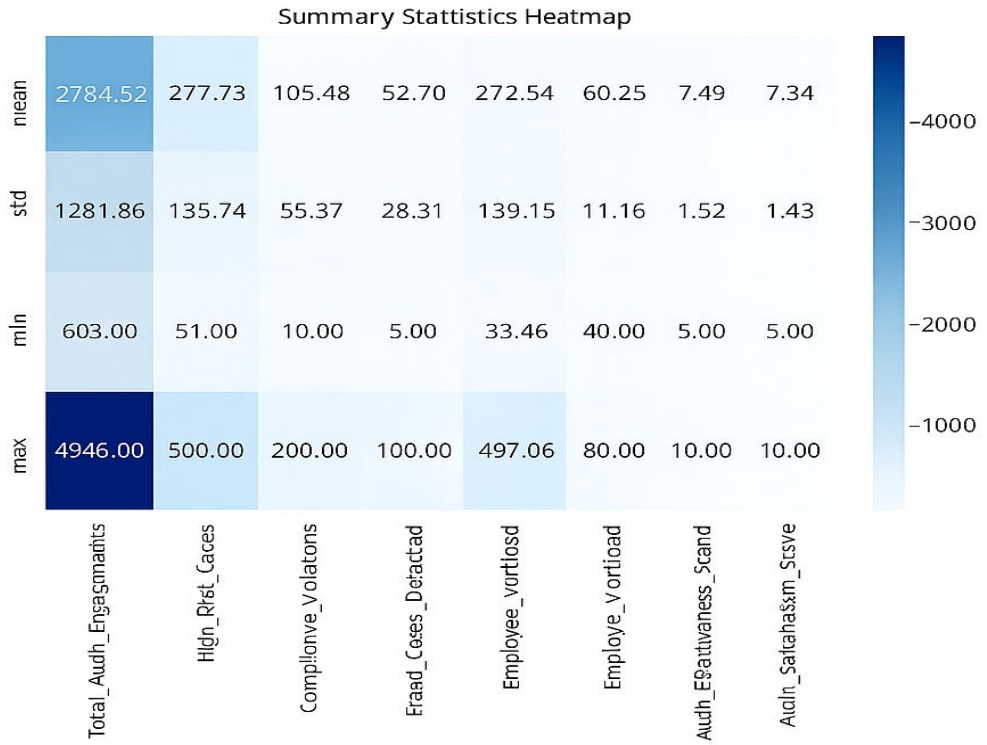
Column names:

- Year
- Firm Name
- Total Audit Engagement
- High-Risk Cases
- Compliance violations
- Fraud Cases Detected
- Industry Affected
- Total Revenue Impact
- AI Used for auditing
- Employee Workload
- Audit effectiveness score
- Client Satisfaction Score

Summary Statistics Heatmap

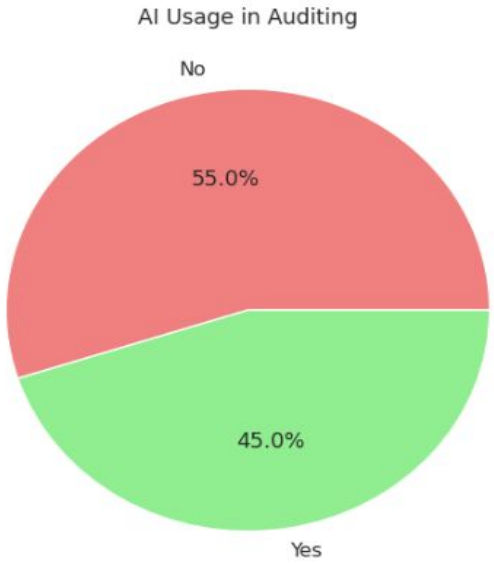
WHAT THIS SHOWS:

- A heatmap of descriptive statistics for key audit and risk-related variables.
- High_Risk_Cases averaged around 277.73, while Fraud_Cases_Detected had a lower mean of 52.70.
- Compliance_Violations showed considerable variation, with values ranging from 10 to 200.
- Total_Revenue_Impact, which reflects financial loss or impact due to risk, ranged from 33.46 to 497.06, with a mean of 272.54.

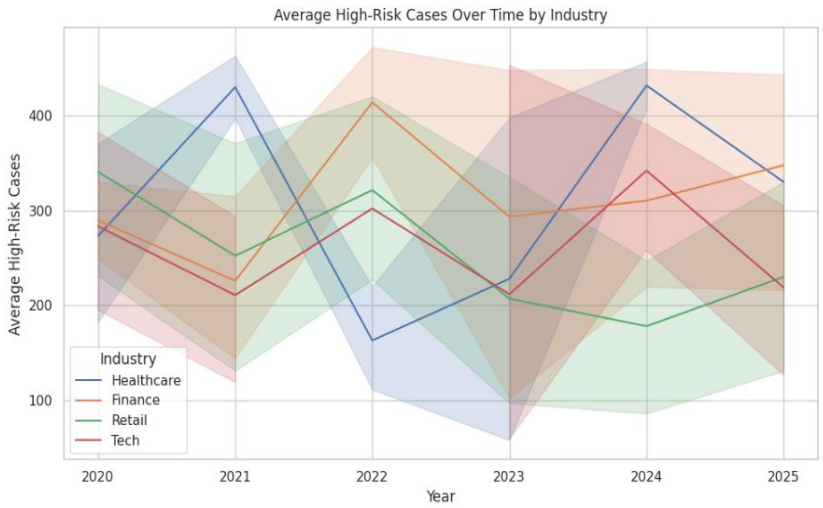


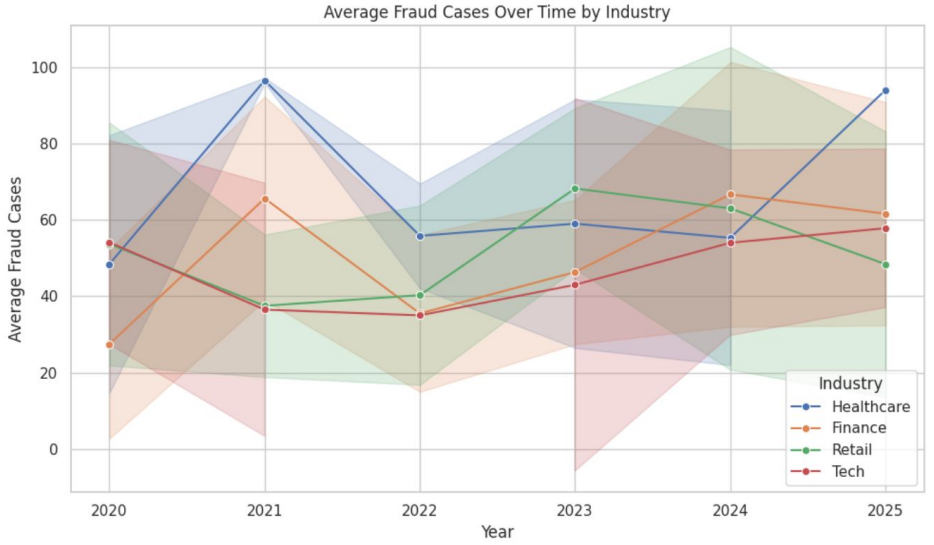
Exploring distribution counts

- The pie chart shows the data is a near even split between audits done with AI usage and not having AI assistance.



Line Chart: Fraud and High-Risk Cases between 2020 and 2025





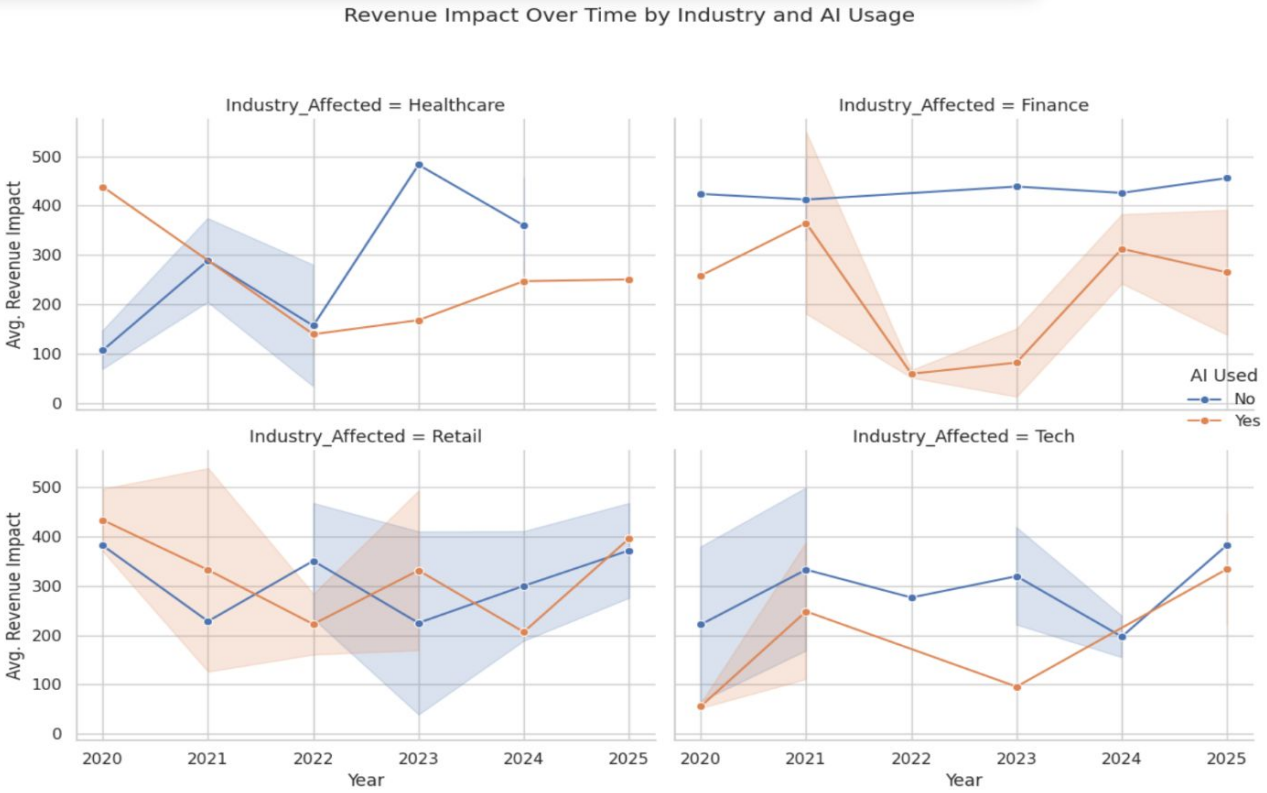
Fraud and high-risk cases show **noticeable fluctuations across industries**, highlighting volatility in audit outcomes.

Retail and Healthcare consistently exhibit **higher levels of fraud and risk**, suggesting the need for enhanced oversight.

Technology and Finance display more stable patterns, though occasional spikes indicate emerging vulnerabilities

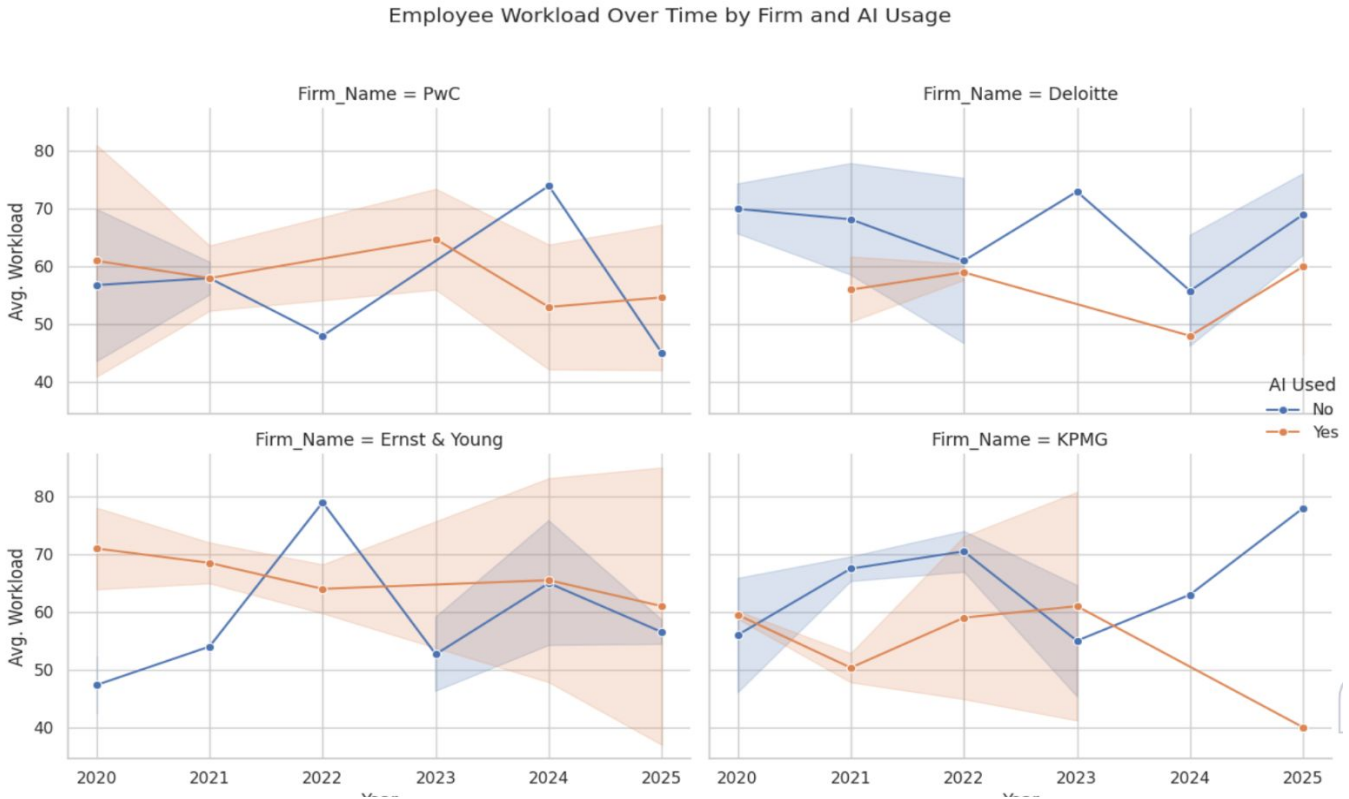
Revenue Impact Over Time by Industry and AI Usage

- **AI usage helped stabilize revenue outcomes**, especially in **Healthcare** and **Retail** where volatility was high.
- In **Finance**, AI adoption supported a **gradual recovery** after a dip in 2022.
- **Tech firms using AI** experienced **reduced volatility** in later years, suggesting better cost control and efficiency.



Employee Workload Over Time by Firm and AI Usage

- **KPMG** and **Ernst & Young** show the **largest reductions in employee workload** with AI adoption.
- **Firms without AI** tend to show **more variability or rising workloads**, especially in **Deloitte** and **PwC**.
- Overall, AI usage appears to help **reduce and stabilize employee workload** over time.

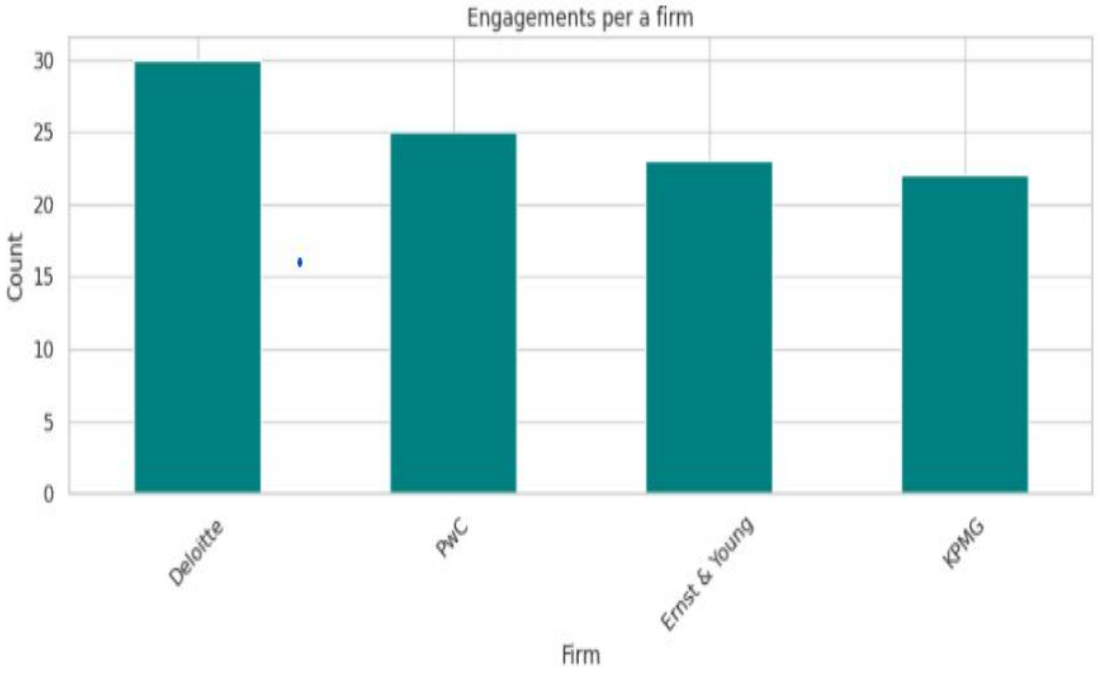


Audit Effectiveness Score by AI Usage (per Firm)

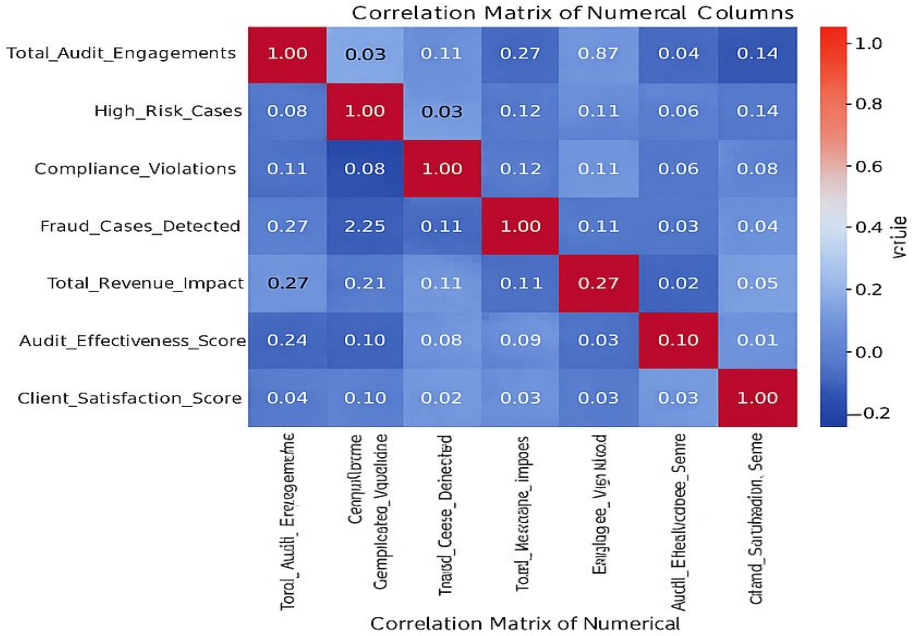
- **AI adoption leads to higher and more stable audit effectiveness scores**, especially in **Ernst & Young** and **PwC**.
- **Firms not using AI** experience more **fluctuation** and generally lower performance.
- **KPMG without AI** shows the sharpest decline, highlighting the risk of non-AI workflows.

Bar Chart: Engagement Distribution per Firm

- Deloitte led with 30 audit engagements,
- Followed by PwC (25),
- Ernst & Young (23),
- KPMG (22). While Deloitte handled slightly more cases, the overall distribution is fairly balanced.



Correlation Matrix of Key Audit Metrics



Key Insights:

- Fraud Cases Detected has the strongest correlation with Total Audit Engagements (r = 0.27), suggesting more audits tend to reveal more fraud.
- Compliance Violations show weak correlations across most variables, with only a mild link to revenue impact (r = 1.
- Employee Workload is negatively correlated with Audit Effectiveness (r = – 0.07), indicating overwork may reduce quality.
- Client Satisfaction slightly improves with Audit Effectiveness (r = 0.10), but decreases with higher audit volumes (r = –0.14).

Data Source: big4_aidit dataset (2020–2025)

Interesting/ Surprising Results

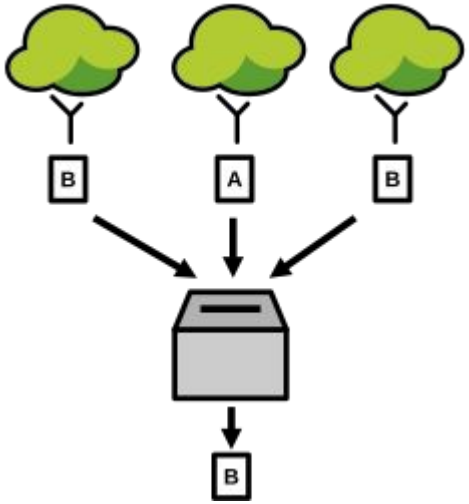
Interesting Results and Patterns

- 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.

Summary of Methods to Solve the Problem

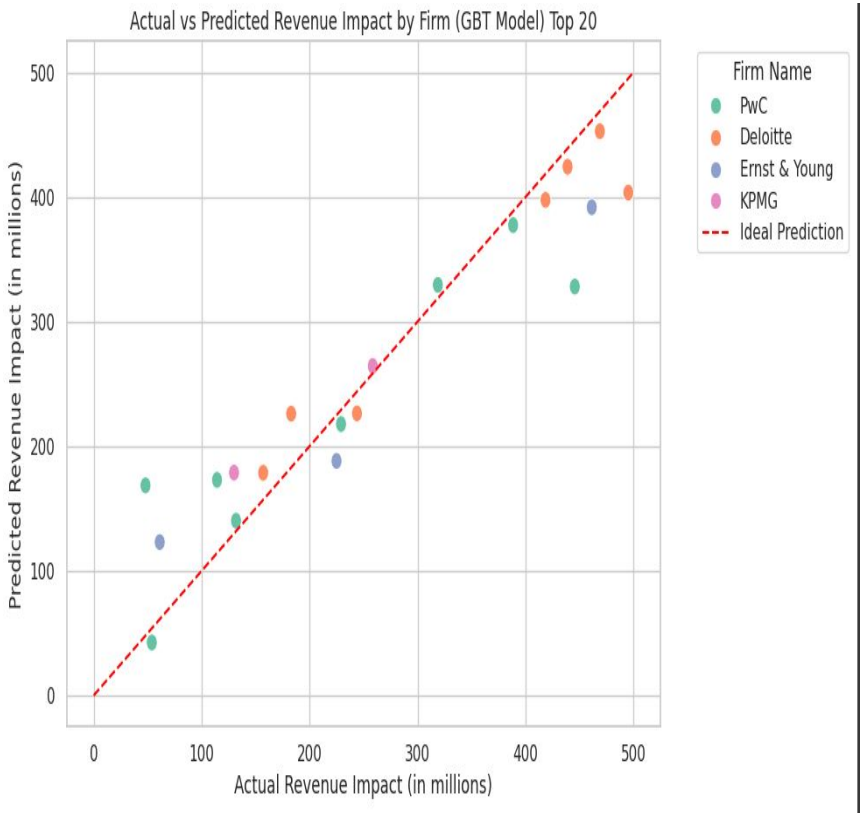
Predictive Modeling:

- **Tree-based models**: i.e *Random Forest*
- **Performance metrics**: *i.e R-squared, MSE, Cross-validation*
- **Key features:**
  -Risk Rate (High-risk cases per audit engagement).
  -Fraud Rate (Fraud cases per violations).
  -**Compliance Per Engagement** (Regulatory adherence per audit).

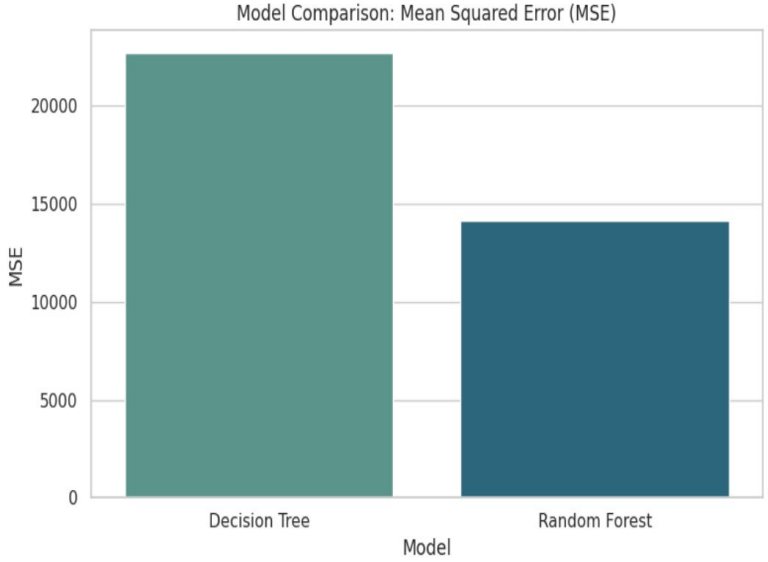


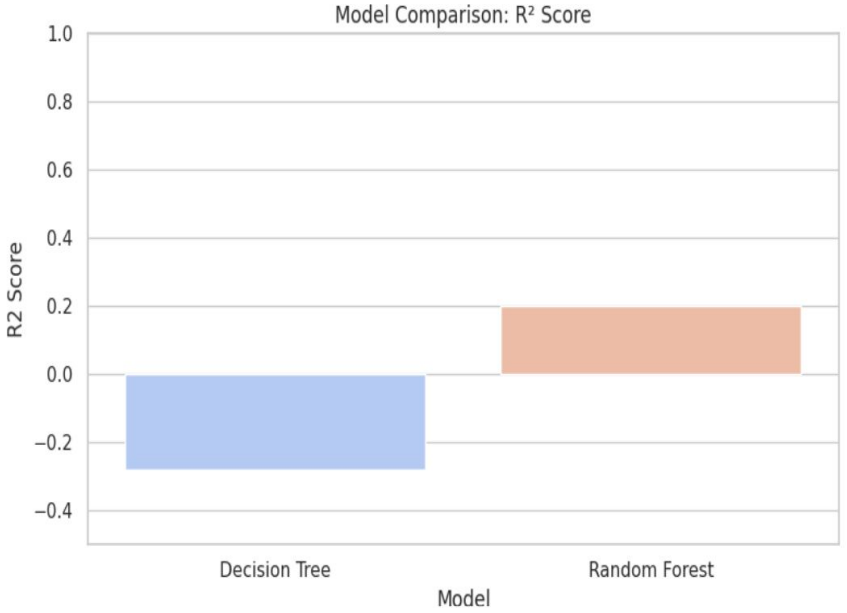
Can financial losses from audit-related compliance failures be predicted?

- A Gradient Boosted Tree model was used to estimate financial loss
- Root Mean Squared Error (RMSE): 66.13 million
- Predictions were made using audit-related features such as: Firm, Industry, AI usage, Workload, Violations, and Fraud Cases
- Most points fall near the ideal line, showing strong prediction accuracy
- No clear over- or under-prediction bias across firms
- Model results are consistent across PwC, Deloitte, EY, and KPMG
- Indicates that audit-related financial losses can be reliably predicted



Decision Tree v. Random Forest for predicting High-Risk Cases model comparison





Decision Tree underperformed (**MSE: 22,729**, **R^2: –0.28**)

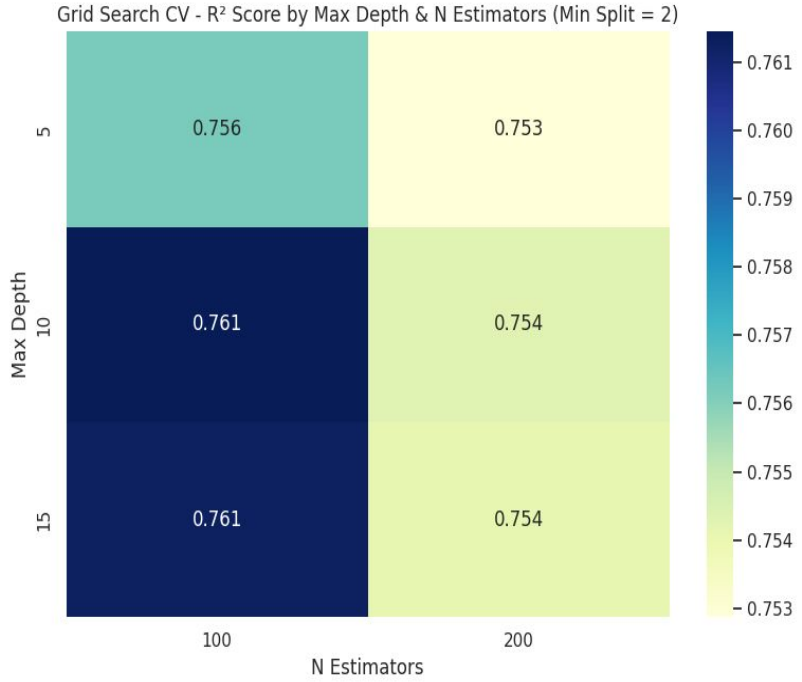
Random Forest has lower error (**MSE: 14,155**) and higher accuracy (**R^2: 0.20**)

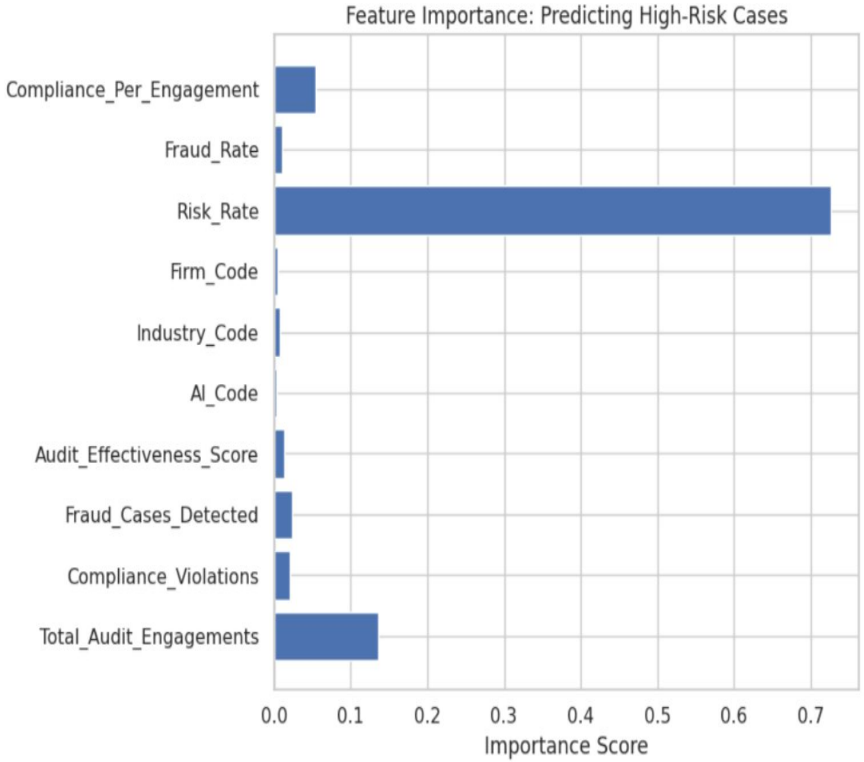
Random Forest + Grid Search Optimization for predicting high risk cases

Model Performance (GridSearchCV)

- **Best Cross-Validated R^2:** 0.76
- **Test R^2 Score:** 0.71
- **Test MSE:** 5,104
- **Optimal Parameters:**
  max_depth=10, min_samples_split=2, n_estimators=100

Interpretation: The model can **accurately predict about 71% of the patterns** that lead to high-risk audit cases. It means the predictions are pretty close to actual outcomes and a major improvement from earlier models.





Top Predictive Features

According to feature importance analysis:

- **Most Influential:**
  - Risk Rate
  - Total Audit Engagements
  - Compliance Per Engagement
- **Lower Impact:**
  - Firm Code, Industry Code, AI Code

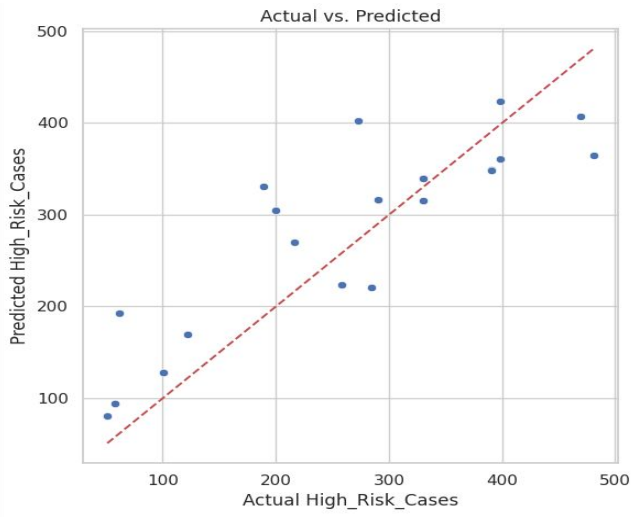
Interpretation: Predictive features align logically — higher audits, fraud, and violations often signal higher audit risk, while categorical identifiers added minimal value.

Actual vs. Predicted – Scatter Plot Analysis

Prediction Accuracy: High-Risk Audit Cases

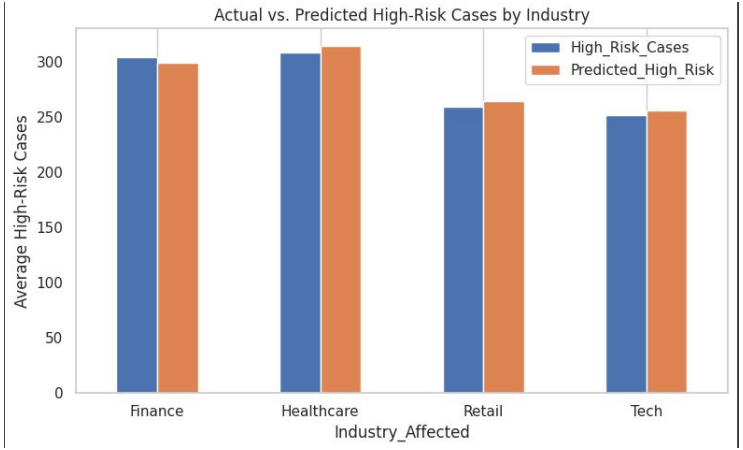
- The scatter plot compares actual vs. predicted high-risk cases.
- Most predictions closely follow the ideal (red dashed) line.
- Slight deviations at higher values, but overall trend is strong.

Insight: The model performs reliably in the 100–400 case range, capturing the underlying relationships well.



Predicted vs. Actual High-Risk Cases by Industry

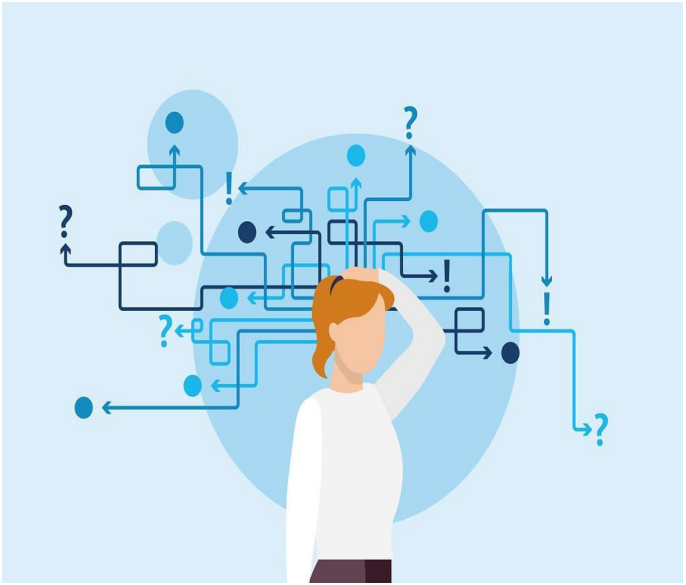
Industry_Affected	High_Risk_Cases	Predicted_High_Risk
Finance	304.250000	299.054583
Healthcare	308.125000	314.278681
Retail	259.259259	264.007407
Tech	251.482759	255.615575



- Predictions are **within ±6 cases** for all industries.
- The model is **well-calibrated** and generalizes well across sectors.

Problems Encountered

- 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 advanced approach was needed , incorporating tree-based models and targeted feature engineering—was necessary to improve performance.



Summary of Results

Overall, the project was successful in building a model to predict high-risk audit cases.

The model also helped estimate potential financial loss per firm.

Key patterns were identified, offering insights into how internal audits (IA) have influenced outcomes , including maximizing audit effectiveness scores, reducing fraud cases, and uncovering risk factors across industries.

Conclusion

- **Data science techniques** enabled strong forecasting of high-risk audit cases, with the **Random Forest model** excelling at pinpointing risk areas.
- **AI adoption** improved audit stability, effectiveness, and reduced employee workload—enhancing both **audit quality and team well-being**.
- Fraud prediction remained **challenging** due to missing behavioral indicators, but key risk-related patterns were still identifiable.
- **Ratio-based features** (e.g., Risk Rate, Compliance Per Engagement) significantly improved model performance by adding context beyond raw data.
- **Predictive analytics**, when combined with industry-specific risk patterns, can help firms proactively manage compliance and financial exposure.

Thank you!