Professor Sreyasee Das Bhattacharjee
May 5, 2020

FINANCIAL STATEMENT FRAUD DETECTION

Two Guys

TEAM MEMBERS
Sangrok Lee
Shardul Rane

INTRODUCTION

Recently, there are many multinational corporations adopt artificial intelligence (AI) in their businesses. It improves analysis (increase accuracy and decrease human error) and increases the competitive advantage to the management. Bloomberg Tax announced that the big four accounting firms, Deloitte, Ernst & Young (EY), KPMG, and PricewaterhouseCoopers (PwC), invested a billion dollars in AI and data technologies. We believe that the continuous development of AI will prevent fraud or material misstatement in real-time and reduce the workload for internal auditors.

Fraud is the intentional manipulation of financial statements such as balance sheets, income statements, and cash flow statements to create a misrepresentation of business performance. For example, Enron was one of the largest companies in the United States and also known for the world's biggest fraud scandal in the 2000s. A material misstatement is the fabricated financial statements that impact on board members and shareholders' decision-making. For example, Enron's off-the-books accounting practices, the shareholders lost their investments caused by stock market bubbles.

PROBLEM STATEMENT

Businesses in the financial sector invested in AI to prevent frauds such as financial statements fraud, credit card fraud, and money laundering. There are various fraud detection models such as anti-money laundering false positive and negative that have put to practical use.

We decided to create a financial statement fraud detection model but limited to the travel expenses account (under the income statement). Reimbursement fraudulent results in the employee obtain reimbursement for expenses that did not occur, duplicate reimbursement, higher reimbursement than the actual cost, unauthorized expenses, and bribes to customers. Auditors spend a considerable time to validate each account before the audit report, and likelihood involve a human error. When examining the travel expenses, auditors need to consider the following:

- Dates of business travel: the employee travel date and end date overlap with an upcoming trip.
- Location: the employee travel city that does not associate with work-related.
- Multiple reimbursements: the employee requests reimbursement for the same travel expenses.
- Frequently report the business travel: the employee committed the same scheme on several expense reports.
- Misuse of the business travel budget: the employee abuses the given budget though reimbursed at a higher price than the actual cost.

Our belief in AI will prevent fraud or material misstatement and reduce the workload for internal auditors, and eliminate the manual processes. Our approach to financial statement fraud detection will analyze the transactions, create a hypothesis(es), and test a hypothesis(es) for potential fraud schemes. Lastly, we will evaluate model risks, internal audit's opportunities and threats on AI, and restructure of three lines of defense (risk management). However, we will not consider all the examples shown above because of limited knowledge and time.

METHOD

- *Anomaly detection* is the process of identifying unexpected items or events in the dataset, which differ from the norm. It often applied to unlabeled data. Anomaly detection practiced in fraud, medical (and healthcare), and structural defects.

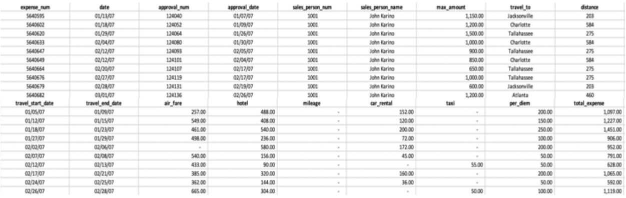
- *Time series anomaly detection* is usually formulated as finding outlier data points relative to some standard or often signal.
- *Ridge Regression* is almost identical to linear regression (sum of square) except the bias added to the model, which a significant drop in variance.
- *One Class Support Vector Machine (SVM)* is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

The Python libraries such as *Pandas, NumPy, Sklearn, Matplotlib, Seaborn, Plotly, Scipy, Keras, Pyod, and et cetera* will helpful in building a financial fraud detection model.

DATASET

First, we acknowledged that the travel expenses account dataset is from the University at Buffalo School of Management, Accounting Department. We will use the dataset for only educational purposes, only in this project, and will not disclose the dataset to others.

The travel expenses account from Tampa Electronics, Inc. It is a manufacturer that distributes the components to other manufacturers. Tampa Electronics has ten full-time employees, salespeople, who travel to various cities to sell components and prove after-sales service. *We modified the dataset; for example, the dataset separated into different Excel sheets, and we merged them into one Excel sheet*.



expense_num	date	approval_num	approval_date	sales_person_num	sales_person_name	max_amount	travel_to	distance
5640595	01/13/07	124040	01/07/07	1001	John Karino	1,150.00	Jacksonville	203
5640602	01/18/07	124052	01/09/07	1001	John Karino	1,200.00	Charlotte	584
5640620	01/29/07	124064	01/26/07	1001	John Karino	1,500.00	Tallahassee	275
5640633	02/04/07	124080	01/30/07	1001	John Karino	1,000.00	Charlotte	584
5640647	02/12/07	124093	02/05/07	1001	John Karino	900.00	Tallahassee	275
5640649	02/12/07	124101	02/04/07	1001	John Karino	850.00	Charlotte	584
5640664	02/20/07	124107	02/17/07	1001	John Karino	650.00	Tallahassee	275
5640676	02/27/07	124119	02/17/07	1001	John Karino	1,000.00	Tallahassee	275
5640679	02/28/07	124131	02/19/07	1001	John Karino	600.00	Jacksonville	203
5640682	03/01/07	124136	02/26/07	1001	John Karino	1,200.00	Atlanta	460

travel_start_date	travel_end_date	air_fare	hotel	mileage	car_rental	taxi	per_diem	total_expense
01/05/07	01/09/07	257.00	488.00	-	152.00	-	200.00	1,097.00
01/12/07	01/15/07	549.00	408.00	-	120.00	-	150.00	1,227.00
01/18/07	01/23/07	461.00	540.00	-	200.00	-	250.00	1,451.00
01/27/07	01/29/07	498.00	236.00	-	72.00	-	100.00	906.00
02/02/07	02/06/07	-	580.00	-	172.00	-	200.00	952.00
02/07/07	02/08/07	540.00	156.00	-	45.00	-	50.00	791.00
02/12/07	02/13/07	433.00	90.00	-	-	55.00	50.00	628.00
02/17/07	02/21/07	385.00	320.00	-	160.00	-	200.00	1,065.00
02/24/07	02/25/07	362.00	144.00	-	36.00	-	50.00	592.00
02/26/07	02/28/07	665.00	304.00	-	-	50.00	100.00	1,119.00

The travel expenses account dataset contains 1,301 rows (observations or transactions) and 18 columns (variables), and the variables consist of:

- *expense_num*: [int], the number of travel reimbursement request form.
- *date*: [date], the date of travel reimbursement request form.
- *approval_num*: [int], the number of approved business travel request form.
- *approval_date*: [date], the date of the approved business travel request form.
- *sales_person_num*: [int], the id number of employees. It contains 1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1009 and 1010.
- *sales_person_name*: [str], the name of employees. It contains Carl Svenson, Gary Hatfield, Ian Botham, Jack Poynter, Jane Duzetsky, Jennifer Johnson, John Karino, Julie Jones, Marie Redwood, and Tom Dolan.
- *max_amount*: [int], the budget for each employee's business travel. He or she cannot go beyond the budget.
- *travel_to*: [str], the location of cities to sell components, and prove after-sales service. It contains Atlanta, Charlotte, Jacksonville, Miami, Mobile, Orlando, and Tallahassee.
- *distance*: [int], the distance (in miles) from Tampa Electronics, Inc to the location of cities to sell components and prove after-sales service. It contains 70, 203, 275, 280, 460, 520 and 584 miles.
- *travel_start_date*: [date], the start of business travel.
- *travel_end_date*: [date], the end of business travel.
- *air_fare*: [int], the amount of an air ticket (round trip) to the location.
- *hotel*: [int], the total amount of hotel expenses.
- *mileage*: [int], the total amount of transportation to the location beside airplane. The employees cannot request reimbursement for both air_fare and mileage.
- *car_rental*: [int], the total amount of rent a car for business purposes.
- *per_diem*: [int], the total amount that employee allows to spend for daily expenditure ($50 per day in 2007, and $56 per day in 2008).
- *total_expense*: [int], the total amount of air_fare, hotel, mileage, car_rental, and per_diem, and the total amount cannot excess the max_amount.

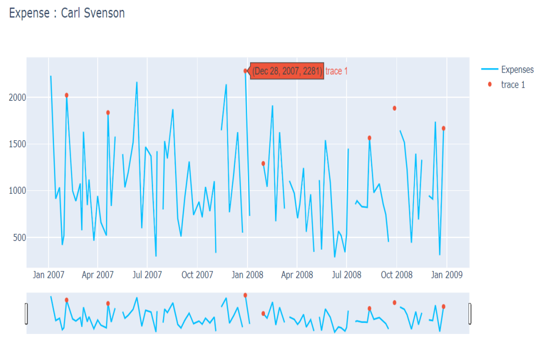
We did data preprocessing before analyzing transactions because the quality of the dataset directly affects the ability of the models to learn; therefore, we must preprocess the travel expenses account dataset. We performed feature scaling though mean normalization (standard scale from *Sklearn*) to Ridge Regression model for estimation of max_amount. Lastly, to perform anomaly detection, we added the new variable, *days_of_travel*, which the number of days between travel_start_date and travel_end_date.

MODEL AND RISK

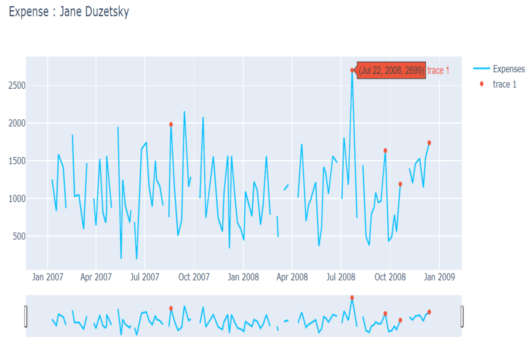
I. Reimbursement Fraud Detection

We created time-series anomaly detection based on each employee's total expenses, which include air_fare, hotel, mileage, car_rental, and per_diem on business travel. We created the model to detect where the employee request reimbursement at a higher price than the actual cost, or request multiple reimbursements.

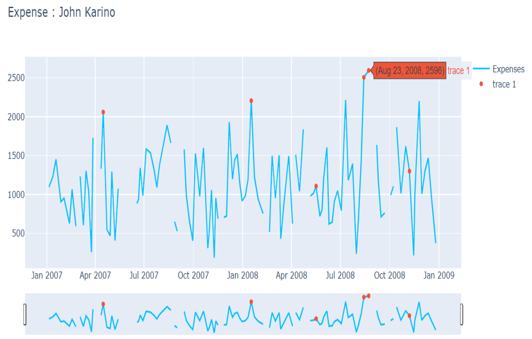
We used a one-class support vector machine (SVM) algorithm that learns the boundaries of points and able to classify any points that lie outside the boundary (outliers). We visualized each employee's total expenses through the graph that a one-class classifier has highlighted the business travel date.



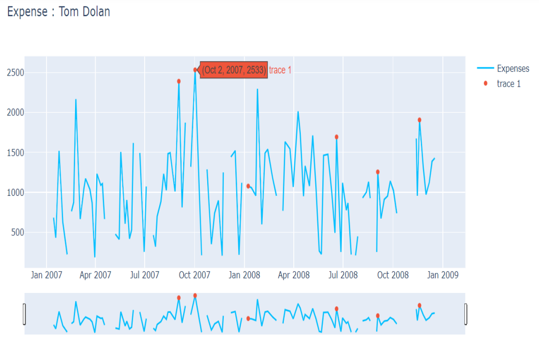
It warns the auditors that on December 28, 2007, Carl Svenson's total expenses were significantly high compare to other business trips.



It warns the auditors that on July 22, 2008, Jane Duzetsky's total expenses were significantly high compare to other business trips.



It warns the auditors that on August 23, 2008, John Karino's total expenses were significantly high compare to other business trips.



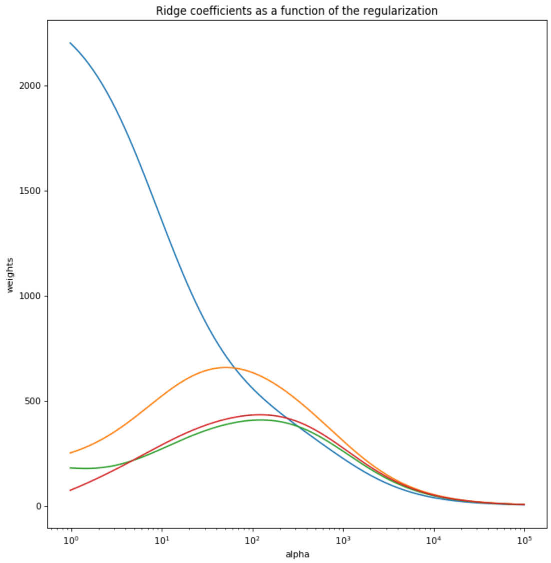
It warns the auditors that on October 2, 2007, Tom Dolan's total expenses were significantly high compare to other business trips.

The time-series anomaly detection model helps the auditors to examine each employee's reimbursement more efficiently than before. However, the outlier does not mean that the employee committed potential fraud or material misstatement, there are reasons such as dinner with executive members in the business, or external factors such as temporary increase price due to conflict with union officials.

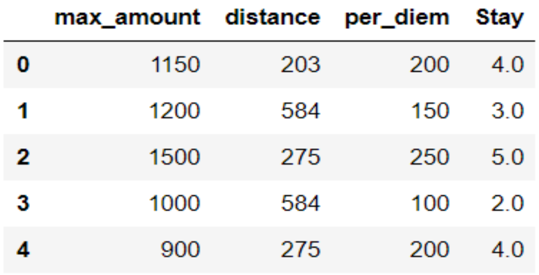
The risk of time-series anomaly detection model is when the characteristics of the signal have changed dramatically, and the model does not work well. For example, when an employee often travels to Charlotte, but he or she changed the location of business travel to Orlando, the model will start to fail due to the high fluctuation of the consumer's price.

II. Business Travel Budget Model

Auditors spend a considerable time to validate each account before the audit report but also require them to share the recommendation on how to improve accounts. The employee might abuse the given budget through reimbursed at a higher price than the actual cost. We created ridge regression based on the budget for each employee's business travel. We created the model that estimates the maximum amount for future business travel.



It shows the change in the regularization weights for four features (includes max_amount, distance, per_diem, and days_of_travel) used in the prediction. We noticed that ridge regression did not introduce any sparsity in the weights.



	max_amount	distance	per_diem	Stay
0	1150	203	200	4.0
1	1200	584	150	3.0
2	1500	275	250	5.0
3	1000	584	100	2.0
4	900	275	200	4.0

We used the training data for estimating the maximum amount for future business travel, but we need more data for better training. The risk of the ridge regression model is heavily influenced by external factors such as the state economy. For example, inflation happens due to an increase in wages, profit push inflation, higher taxes, or printing more money.

AUDITOR'S OPPORTUNITY AND RISK ON AI

Our model, the financial statement fraud detection, will reduce workload for internal auditors (significant decrease in human errors) and eliminate the manual processes on validating on each account. However, internal auditors should evaluate the opportunities and risk of AI, and consider the internal auditor's role in AI.

I. Internal Auditor's Opportunity on AI

- AI will eliminate human errors in the transaction validation, and replace time-intensive with time-effective, but more important; it will reduce the cost of labor.
- AI prediction accuracy is more dependable than human prediction accuracy.

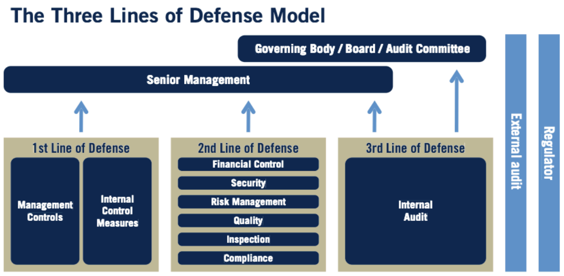
II. Internal Auditor's Risk on AI

- AI created by data scientist, which includes bias and human logic errors because of no foundation of accounting.
- If the business does not invest time and money, the model will be outdated, which mentioned the risks on each model from the previous section.
- Failure of the model will result in ethically questionable results.

III. Internal Auditor's Role in AI

- Internal auditors should maximize the opportunities of AI but manage the risks followed by AI auditing framework: AI governance, data architecture and infrastructure, data quality, the measuring performance of AI, the human factor, and the black box factor.
- Internal auditors should include AI risk assessment, and consider whether to include AI risk in the audit plan.
- Internal auditors should ensure the moral and ethical issues of AI.

RESTRUCTURE OF THREE LINES OF DEFENSE



The three lines of defense are to understand the system of internal control, and risk management should not regard as an automatic guarantee of success. All three lines need to work effectively with each other and with the audit committee to create the right conditions. Each line of defense consists of the following:

- The first line of defense: management controls and internal control measures.
- The second line of defense: financial controller, security, risk management, quality, inspection, and compliance.
- The third line of defense: internal audit.

The effective risk management (three lines of defense) has saved countless businesses from irreparable damages in the past, in which businesses invest time and money to manage the risks.

The implementation of financial statement fraud detection will require the restructure on the third line of defense. The collaboration of our model and internal auditors will lead to an improvement in assurance (on the effectiveness of risk management) and the highest level of independence within the business.

DISTRIBUTION OF PROJECT

We equally distribute the work based on the background of education and experiences. For example, Sangrok has an accounting and finance background, and Shardul has a computer science background, which Sangrok focused on Business Travel Budget model and report, especially the integration of AI and internal audit. Shardul focused on Reimbursement Fraud Detection, each model risks based on algorithms, and UI for presentation. Lastly, we wrote a report through Google Docs. We believe that we work effectively and efficiently during the pandemic.

REFERENCES

"Big Four Invest Billions in Tech, Reshaping Their Identities." *Bloomberg BNA News*, 2 Jan. 2020, news.bloombergtax.com/financial-accounting/big-four-invest-billions-in-tech-reshaping-their-identities.

CA, Steve Bruce. "Internal Audit: Three Lines of Defence Model Explained." Icas.com, 4 Oct. 2019, www.icas.com/professional-resources/audit-and-assurance/internal-audit/internal-audit-three-lines-of-defence-model-explained.

Gandhi, Rohith. "Support Vector Machine - Introduction to Machine Learning Algorithms." *Medium*, Towards Data Science, 5 July 2018, towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

"Governance of Risk: Three Lines of Defence." IIA, www.iia.org.uk/resources/audit-committees/governance-of-risk-three-lines-of-defence/.

Li, Susan. "Anomaly Detection for Dummies." *Medium*, Towards Data Science, 2 July 2019, towardsdatascience.com/anomaly-detection-for-dummies-15f148e559c1?gi=d72887c5c965.

Maklin, Cory. "Machine Learning Algorithms Part 11: Ridge Regression, Lasso Regression And Elastic-Net Regression." Medium, Medium, 31 Dec. 2018, medium.com/@corymaklin/machine-learning-algorithms-part-11-ridge-regression-7d5861c2bc76.

Nickolas, Steven. "What Is Accounting Fraud?" *Investopedia*, Investopedia, 28 Feb. 2020, www.investopedia.com/ask/answers/032715/what-accounting-fraud.asp.

Pandey, Pranjal. "Data Preprocessing : Concepts." *Medium*, Towards Data Science, 25 Nov. 2019, towardsdatascience.com/data-preprocessing-concepts-fa946d11c825.

Segal, Troy. "Enron Scandal: The Fall of a Wall Street Darling." *Investopedia*, Investopedia, 29 Jan. 2020, www.investopedia.com/updates/enron-scandal-summary/.

"The Institute of Internal Auditors." *THE THREE LINES OF DEFENSE IN EFFECTIVE RISK MANAGEMENT AND CONTROL*, The Institute of Internal Auditors North America, Jan. 2013, na.theiia.org/standards-guidance/Public Documents/PP The Three Lines of Defense in Effective Risk Management and Control.pdf.

Tiunov, Pavel. "Time Series Anomaly Detection Algorithms." *Medium*, Stats and Bots, 7 Mar. 2019, blog.statsbot.co/time-series-anomaly-detection-algorithms-1cef5519aef2.