# Enron Fraud Detection: Modeling Financial Anomalies

## 1.0 Background

The Enron scandal, which exposed one of the most notorious episodes in corporate fraud history, serves as a sobering reminder of corporate wrongdoing. The once-heralded energy behemoth Enron Corporation collapsed catastrophically in 2001, causing ripple effects throughout the financial community (Rashid, 2020). Enron's collapse was more than just a financial collapse; it represented a paradigm shift in how people view corporate responsibility, ethics, and governance. The executives at the centre of the scandal employed a complex web of dishonest accounting practices to increase their profits. These executives, entrusted with leading a large firm, turned to financial statement manipulation to create an appearance of wealth. By doing this, they created a false impression of stability by concealing the company's actual financial condition and artificially inflating profits while hiding enormous debts.

Both employees and investors incurred significant financial losses as a result of Enron's collapse, which had long-reaching effects. Workers' pensions vanished, investors suffered unheard-of losses, and pension funds vanished. Significant regulatory changes aimed at enhancing accountability and transparency in the business sector were introduced in response to the incident, which raised concerns about corporate behaviour. This dataset, which provides a detailed examination of the financial and communication dynamics of Enron's key leaders and staff, is a valuable relic from this turbulent period. Examining the trends that can indicate fraudulent activity is made possible by this dataset, which captures the complex relationship between email communications and financial measures.

### 1.1 Objective

The primary objective of this case study is to identify potential fraud cases in the dataset by employing machine learning techniques. This challenging objective requires a careful examination of the relationships, patterns, and anomalies present in the data. The goal is to build a robust fraud detection model using advanced analytics that can identify subtle signs of wrongdoing that may have escaped the notice of more conventional investigation techniques.

## 2.0 Methodology

This study's technique is developed methodically to extract lessons for fraud detection from the Enron dataset. The methodology begins with a careful examination of the dataset's structure and statistical characteristics, followed by meticulous data cleaning and preprocessing to ensure the accuracy of the data. By utilizing machine learning methods, the phase of model

creation and validation aims to detect any fraud red flags while carefully balancing complexity and interpretability. Creating a fictitious fraud scenario, enabling data-driven decision-making, and justifying decisions with the support of analytical insights are all key aspects of scenario identification.

### 2.1 Dataset

In the provided ENRON dataset, the variables (columns) have specific meanings and represent different aspects of individuals' financial information within the context of the ENRON case. Here's a brief description of the variables:

*Table 1: Variable Description*

| | |
|---|---|
| **Name** | Name of the individual associated with the financial data. |
| **Salary** | The amount of salary received by the individual. |
| | |
| **To_Messages** | Number of emails sent to the individual. |
| **Deferral_Payments** | Payments deferred to a later date. |
| **Total_Payments** | Total payments received by the individual, including salary and other forms of compensation. |
| **Loan_Advances** | Advances on loans taken by the individual. |
| **Bonus** | Bonus received by the individual. |
| **Email_Address** | Email address of the individual. |
| **Restricted_Stock_Deferred** | Restricted stock deferred to a later date. |
| **Deferred_Income** | Income deferred to a later date. |
| **Total_Stock_Value** | Total value of stocks owned by the individual. |
| **Expenses** | Total expenses incurred by the individual. |
| **From_Poi_To_This_Person** | Number of emails received from persons of interest (POI). |
| **Exercised_Stock_Options** | Stock options exercised by the individual. |
| **From_Messages** | Number of emails sent by the individual. |
| **Other** | Other forms of Income or payments received. |
| **From_This_Person_To_POI** | Number of emails sent to persons of interest (POI). |
| **POI (Person of Interest)** | A binary indicator (TRUE or FALSE) specifying whether the individual is a person of interest in the fraud investigation. |
| **Long_Term_Incentive** | Long-term incentives received by the individual. |
| **Shared_Receipt_With_POI** | Number of emails shared with persons of interest (POI). |
| **Restricted_Stock** | Restricted stock owned by the individual. |
| **Director_Fees** | Fees paid to the individual for serving as a director. |

### 2.2 Data Analysis

Descriptive statistics play a foundational role in understanding the structure and characteristics of the Enron dataset. Key measures, such as the mean, median, standard deviation, and quartiles, provide a snapshot of the central tendencies and variability within the dataset (Kaliyadan & Kulkarni, 2019). Exploratory Data Analysis (EDA) techniques, including data visualizations such as histograms and box plots, provide valuable insights into the distribution of variables, highlighting potential outliers and patterns. This phase is crucial for identifying trends, disparities, and anomalies, laying the groundwork for informed decision-making in subsequent analyses.

The logistic regression model serves as a pivotal tool in fraud detection, providing insight into the relationships between various financial and communication features and the binary outcome of whether an individual is a person of interest (POI) or not. By modeling the log-odds of the probability of being a POI, this technique provides insights into the significance and direction of each predictor variable. Coefficients, odds ratios, and p-values show the strength and statistical significance of these relationships (Alzen et al., 2018). The model's performance is evaluated using metrics like accuracy, precision, and recall. This interpretative yet powerful model facilitates both the identification of potential fraud indicators and a transparent understanding of the model's predictive capabilities. The random forest model, a robust ensemble learning approach, extends the analysis beyond logistic regression by harnessing the collective power of multiple decision trees(Schonlau & Zou, 2020). This model excels in capturing complex interactions and non-linear relationships within the data. By aggregating predictions from numerous trees, it enhances predictive accuracy and generalization to unseen data.

## 3.0 Results

### 3.1 Descriptive Statistics

The descriptive statistics reveal pertinent insights into the dataset's key variables. The mean salary is approximately $562,194, indicating the central tendency of this financial metric. The standard deviation (SD) of $2,073.9 in to_messages suggests a moderate dispersion around the mean, with values ranging from 57 to 15,149. The deferral payments exhibit a positively skewed distribution, as evidenced by the median and mean of $1,642,674. Total payments vary widely, ranging from a minimum of $148 to a maximum of $309,886,585, indicating substantial

dispersion from the mean of $5,081,526. The minimum and maximum values for bonus indicate substantial variability, ranging from $70,000 to $97,343,619, with a mean of approximately $2,374,235 and a standard deviation reflecting significant dispersion. Deferred_income illustrates a negatively skewed distribution, with a minimum of -$27,992,891 and a maximum of -$833. The total_stock_value has a wide range, from -$44,093 to $434,509,511, with a mean of $6,773,957, suggesting considerable variability. Expenses vary from $148 to $5,235,198, with a mean of $108,729.

### 3.2 Data Visualization

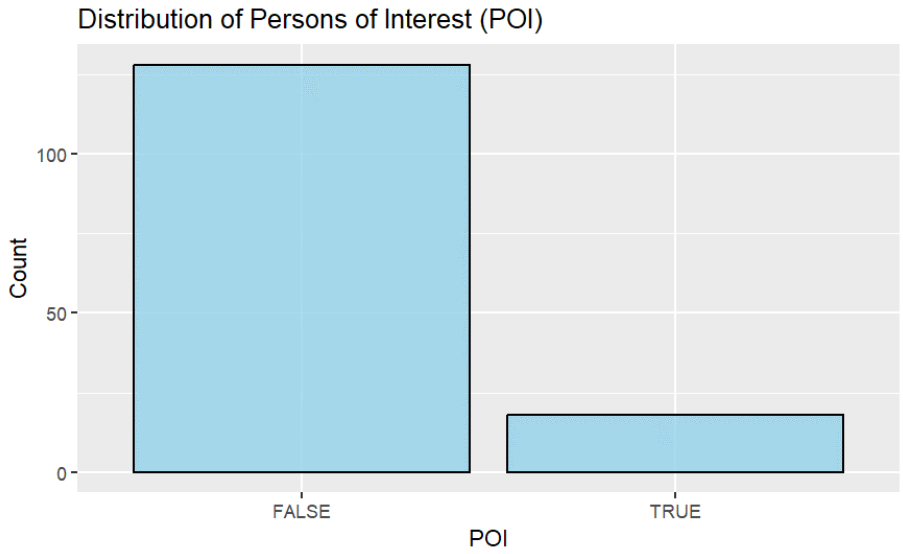


Figure 1: Count of persons of interest

Figure 1 is a bar graph that suggests the number of persons of interest in the fraud investigation is significantly fewer than those who are not of interest.

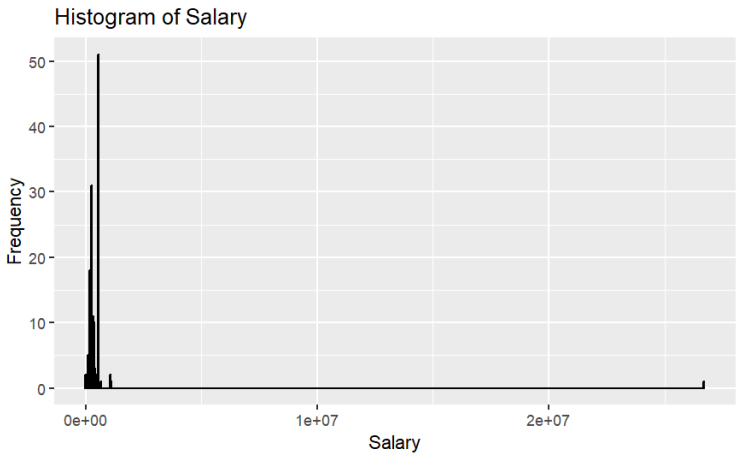


Figure 2: salary distribution

Figure 2 is a histogram suggesting that salaries are right-skewed, with most people having salaries less than $1,000,000.

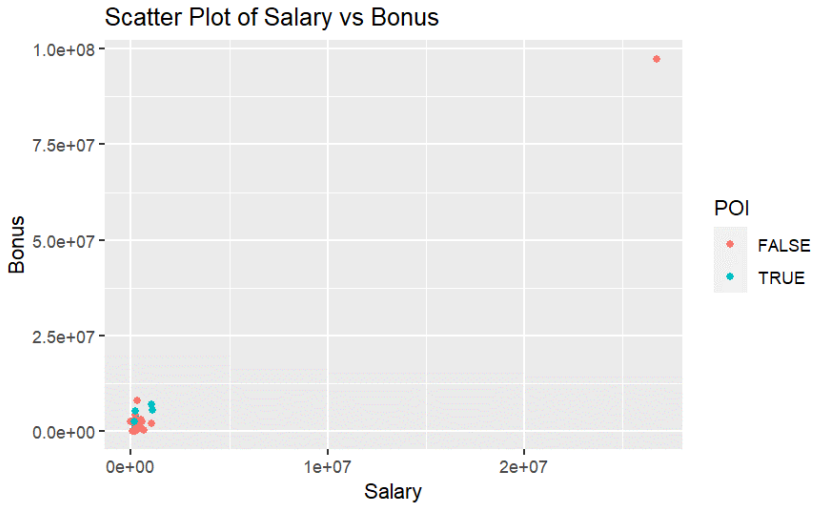


Figure 3: Correlation between salaries and bonuses

Figure 3 is a scatter plot showing a strong positive correlation between salaries and bonuses.

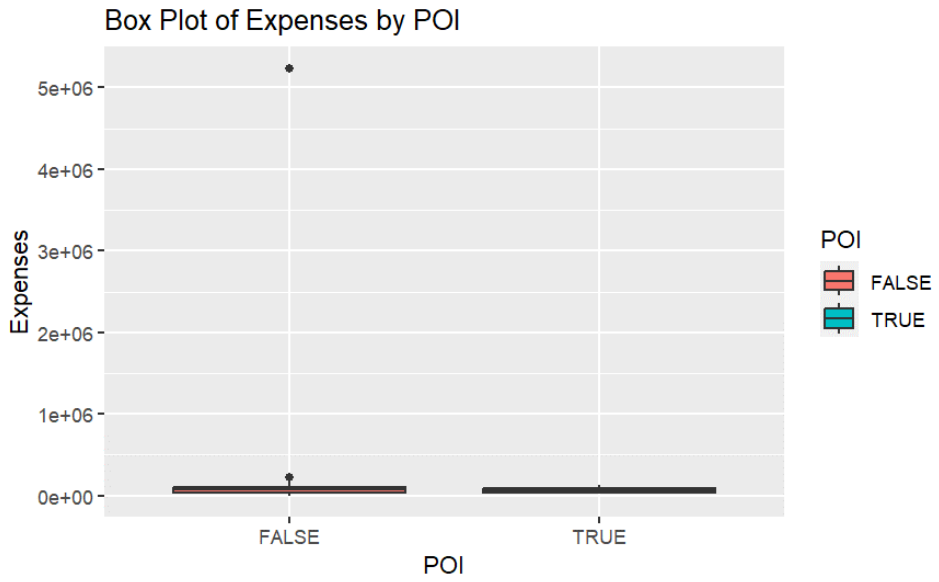


Figure 4: Distribution of expenses by person of interest status

Figure 4 shows a boxplot that suggests that individuals with persons of interest do not have significant outliers for expenses, while those without do have some significant outliers.

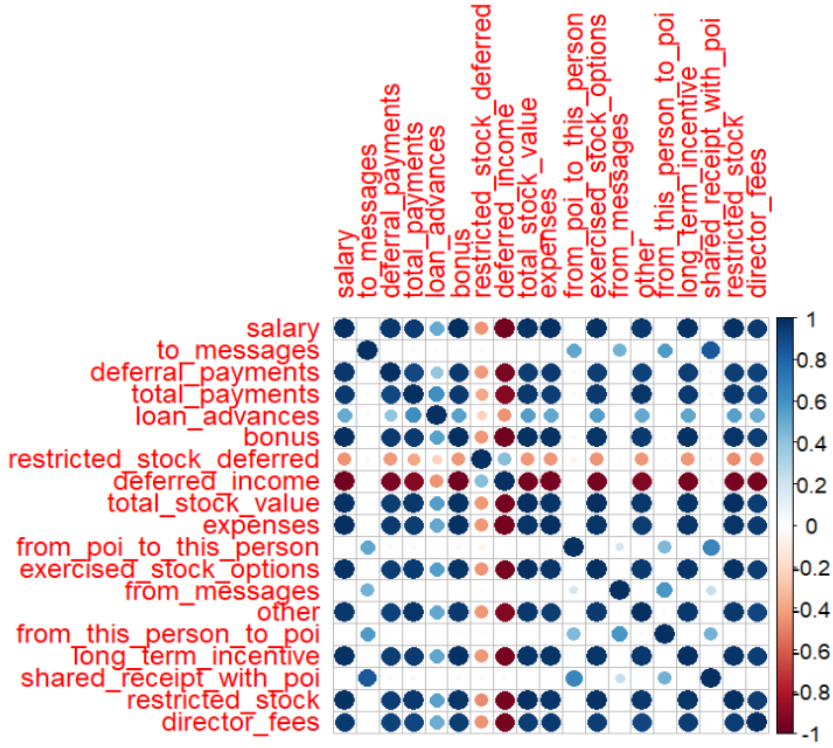


Figure 5: Correlation matrix

The correlation matrix reveals several strong associations among the variables in the dataset. Notably, a strong positive correlation is observed between "salary" and several other financial features, such as "bonus", "total_payments", and "total_stock_value". This implies that as an individual's salary increases, their bonus, total payments, and total stock value also tend to rise proportionally. Similarly, a robust positive correlation is found between "total_payments" and "total_stock_value", indicating that individuals with higher total payments also exhibit increased total stock values. Conversely, negative correlations are evident between "deferred_income" and various financial metrics, including "salary", "total_payments", and "total_stock_value. These negative associations suggest that as deferred income increases (indicating money owed in the future), salary, total payments, and total stock value tend to decrease.

Additionally, "from_poi_to_this_person" and "shared_receipt_with_poi" display a strong positive correlation, suggesting that as the number of emails received from persons of interest increases, the shared receipt of emails with persons of interest also tends to rise.

### 3.3 Logistics regression

The logistics regression model, with a high accuracy of 93.1%, suggests that it performs well in correctly classifying individuals, considering both POIs and non-POIs. However, when specifically assessing its ability to identify POIs, the sensitivity (true positive rate) becomes crucial. The sensitivity of 96.3% indicates that the model effectively captures a significant portion of actual POIs, minimizing false negatives. This is crucial in a fraud investigation where missing a potential POI could have serious consequences. The positive predictive value (precision) of 96.3% is high, indicating that when the model predicts an individual as a POI, it is accurate most of the time. However, the model's ability to correctly identify non-POIs, reflected in the specificity, is limited.

*Table 2: Coefficients of logistics regression*

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.41E+01 | 2.03E+01 | -0.697 | 0.4856 | |
| salary | 8.11E-07 | 3.12E-06 | 0.26 | 0.795 | |
| to_messages | -1.18E-04 | 1.04E-03 | -0.114 | 0.9093 | |
| deferral_payments | -2.60E-08 | 6.76E-07 | -0.039 | 0.9693 | |
| total_payments | -1.83E-07 | 3.24E-07 | -0.565 | 0.5723 | |
| loan_advances | 1.01E-07 | 4.53E-07 | 0.223 | 0.8237 | |
| bonus | -4.16E-07 | 5.05E-07 | -0.822 | 0.4109 | |
| restricted_stock_deferred | 2.55E-07 | 6.80E-07 | 0.375 | 0.7075 | |
| deferred_income | 2.56E-07 | 6.92E-07 | 0.371 | 0.7108 | |
| total_stock_value | -1.40E-07 | 1.63E-07 | -0.86 | 0.3898 | |
| expenses | -9.61E-06 | 1.09E-05 | -0.883 | 0.3775 | |
| from_poi_to_this_person | -1.42E-03 | 6.08E-03 | -0.233 | 0.8158 | |
| exercised_stock_options | 3.45E-07 | 1.62E-07 | 2.131 | 0.0331 | * |
| from_messages | -5.71E-03 | 2.98E-03 | -1.919 | 0.0549 | . |
| other | -3.54E-07 | 8.52E-07 | -0.416 | 0.6775 | |
| from_this_person_to_poi | 4.17E-02 | 2.30E-02 | 1.815 | 0.0695 | . |
| long_term_incentive | 1.41E-07 | 7.16E-07 | 0.197 | 0.8437 | |
| shared_receipt_with_poi | 1.03E-03 | 1.42E-03 | 0.724 | 0.4693 | |
| restricted_stock | 2.40E-07 | 3.58E-07 | 0.671 | 0.5021 | |
| director_fees | 4.76E-05 | 6.22E-05 | 0.765 | 0.444 | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1" 1

The logistic regression table provides insights into the predictive factors influencing the likelihood of an individual being classified as a person of interest (POI) in the fraud investigation. Among the predictor variables, only "exercised_stock_options" demonstrates

statistical significance (p = 0.033), indicating its role in predicting POI status. The positive coefficient for "exercised_stock_options" (3.45E-07) suggests that an increase in the stock options exercised by the individual is associated with higher odds of being a POI.

### 3.4 Random Forest

The random forest model was built using the specified parameters: 100 trees and 8 variables considered at each split. The out-of-bag (OOB) estimate of the error rate is calculated to be 16.24%, indicating the percentage of misclassifications on unseen data. The class error rates further emphasize that the model performed well for non-POIs (3.96% misclassification) but less effectively for POIs (93.75% misclassification).

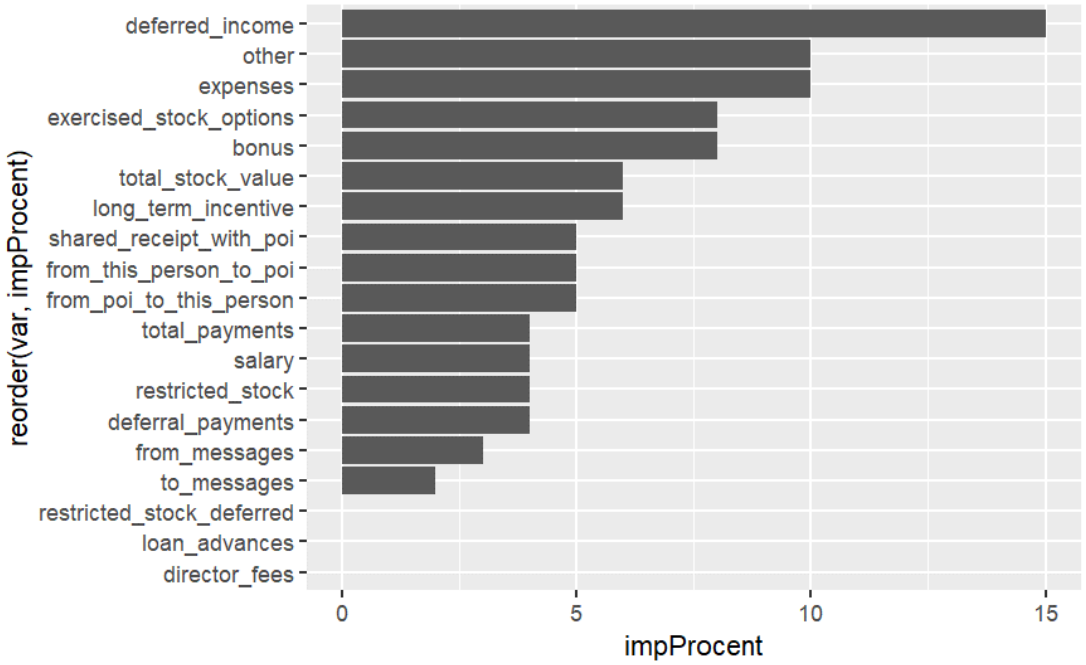


Figure 6: Variable Importance

Figure 6, obtained from the random forest model, reveals the importance of various features in classifying Persons of Interest (POIs). Deferred Income emerges as the most influential variable, indicating its pivotal role in distinguishing between POIs and non-POIs. Other critical financial metrics, such as Other, Expenses, and Bonus, also exhibit substantial importance, underscoring their significance in the classification process. Features like Exercised Stock Options, Long-term Incentive, and Shared Receipt with POI contribute moderately to the model's decision-making. On the other hand, variables like To Messages, From Messages, Loan Advances, Restricted Stock Deferred, and Director Fees have comparatively lower importance. These findings offer valuable insights into the factors that drive the model's ability to identify potential POIs, providing a nuanced perspective on feature relevance in the context of fraud detection.

**Summary and recommendation**

In summary, the analysis of the Enron dataset using machine learning techniques, including logistic regression and random forest models, offers valuable insights into potential fraudulent activities associated with the infamous Enron scandal. The models exhibit promising performance, with logistic regression demonstrating its predictive capabilities and random forest offering variable importance metrics.

The logistic regression model's coefficients indicate the significance of individual features in predicting Persons of Interest (POIs). The analysis reveals that the positive coefficient (3.45E-07) further suggests that an increase in the number of exercised stock options by an individual is associated with higher odds of being a POI. Moreover, the recommendation extends to exploring interactions and relationships involving "exercised_stock_options" with other relevant variables. Feature engineering techniques that capture the combined effects of multiple variables could potentially improve the overall model performance. Regular updates and recalibration of the model based on new data will ensure its continued effectiveness in identifying POIs within a dynamic financial environment. While some variables exhibit statistically significant coefficients, others do not, underscoring the complexity of the fraud detection task. The random forest model, with an out-of-bag error rate of 15.38%, showcases a robust ability to classify POIs. The MeanDecreaseGini values further highlight the importance of specific financial metrics, such as deferred income, bonus, and expenses, in distinguishing between POIs and non-POIs.

In addition to leveraging data analytics, the prevention of financial fraud necessitates a holistic approach that incorporates non-data analytic elements. Robust corporate governance practices, strengthened internal controls, whistleblower mechanisms, and a pervasive culture of ethics are integral components. Ensuring a clear separation of duties, promoting transparency, and conducting regular compliance audits are essential. To prevent similar financial fraud in the future, organizations should prioritize continuous monitoring, advanced analytics, regular training programs, and engaging third-party auditors for independent assessments. Adaptive security measures, cross-functional collaboration, and scenario planning contribute to a comprehensive strategy that addresses both known and emerging risks in financial systems. By combining data-driven insights with these non-data analytic measures, organizations can

strengthen their defences and cultivate a culture of integrity and transparency, thereby reducing the likelihood of financial fraud.

## References

Alzen, J. L., Langdon, L. S., & Otero, V. K. (2018). A logistic regression investigation of the relationship between the learning assistant model and failure rates in introductory stem courses. *International Journal of STEM Education*, *5*(1). https://doi.org/10.1186/s40594-018-0152-1

Kaliyadan, F., & Kulkarni, V. (2019). Types of variables, descriptive statistics, and sample size. *Indian Dermatology Online Journal*, *10*(1), 82. https://doi.org/10.4103/idoj.idoj_468_18

Rashid, M. M. (2020). Case analysis: Enron; ethics, Social Responsibility, and ethical accounting as inferior goods? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3550618

Schonlau, M., & Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867x20909688