Data Mining Final Project

Haokun Zhang, Lu Zhang, Jyun-Yu Cheng

2023/04/21

Abstract

In this project, we try to build a predictive model on how internal control affects a company's audit fee. We explore 5 models and select the most accurate model by unsupervised and supervised learning methods. At last, we find that the random forest model is the most accurate predictive model. We find that a company's revenue, whether the auditor belongs to big 4 firms, and the company's asset are the top 3 important factors affecting the decision making of audit fee.

Introduction:

Companies listed on U.S. stock exchanges are required by law to have their financial statements audited by an independent auditor. This requirement is designed to provide assurance to stakeholders, such as investors, creditors, and regulators, that the company's financial statements are presented fairly and accurately in accordance with accounting standards and can help to build trust and confidence in the company. An audit fee then is necessary to be paid to the independent auditor for their services in conducting the audit of the company's financial statements. The audit fee is typically paid annually and covers the cost of the auditor's time, expertise and resources required to conduct the audit as well as potential legal and reputational risks associated with the audit.

To be specific, (1)Time: Auditors spend a significant amount of time reviewing a company's financial statements, internal controls, and other relevant information. The amount of time required depends on the size and complexity of the company, as well as the scope of the audit. Market capitalization (market cap), scale of assets, scale of revenue, scale of earnings etc. could be useful indicators of complexity of companies. (2) Expertise and resource required: Auditors are highly trained professionals with specialized knowledge in accounting, auditing, and financial reporting. They use this expertise to assess the accuracy and reliability of a company's financial statements, as well as to identify potential risks and areas for improvement. Among auditors in global scale, Deloitte & Touche LLP; PricewaterhouseCoopers LLP; Ernst & Young LLP; and KPMG (as known as the Big 4 firms) are usually considered to be more expertise than non-Big 4 firms due to their extensive global networks, significant presence in multiple countries, heavy investment in training and development programs for their staff as well as great investment in technology and innovation. (3) Potential legal and reputational risk: If a material misstatement is later discovered in the financial statements, the auditors may be held liable for any losses suffered by investors or other stakeholders. Meanwhile, if an auditor's work is called into question or their reputation is damaged, it can be difficult for them to attract new clients or retain existing ones. An adverse opinion on internal control over financial reporting (ICFR) could be an indicator of potential higher legal and reputational risk for auditors.

In this project, we answer the question of what are the key factors to determine audit fee, and select the optimal model to predict it. Audit fees are important both for companies and auditors. For companies, predicting audit fees can help companies compare the cost of their audit with other companies in the same industry to assess their competitive position and identify opportunities for cost savings. For auditors, this

research question could help them have a better audit fees negotiation process. A good reference of audit fees could on one hand help them ensure that their fees are reasonable and proportionate to the work performed, which can help to maintain the integrity of the audit profession, on the other hand, help auditors to assess their competitive position in the market.

Data and data visualization

The original data comes from Audit Analytics®, a provider of independent research and data on public companies' financial reporting, auditing, and regulatory compliance (1). Audit Analytics® includes information on audit fees, auditor changes, internal controls, and other financial reporting and governance data for publicly traded companies in the United States and Canada. We collect internal control.csv of fiscal year 2021 with filter condition of market capitalization greater than $75 million since only companies listed on U.S. stock exchanges with market capitalization greater than $75 million are required to secure an independent audit opinion on internal control over financial reporting(ICFR) (details are provided in appendix). The original data consists of 3253 rows of different public companies listed on U.S. stock exchanges and 38 variables out of which we use 18 variables. Considering that data such as market capitalization, total fees, audit fees as well as earnings vary greatly different and skewed, which can make it difficult to perform certain statistical tests or make accurate predictions, we then introduce box-cox transformations and natural logarithm transformations method to normalize the data. In addition to non-zero processing for 0 values and processing for the transformation of properties of variables, we created 11 new variables in total. The details of the each variable present in the following table.

Table 1: Variable Descriptions

Variable	Description
company	Registrant's current name as filed with the SEC
city	Current business address information
state_code	Current business address information 1
state_name	Current business address information 2
state_region	Current business address information 3
auditor	Name of the auditor who signed the auditor's report on the audit of the financial statements
auditor_key	Unique numeric identifier assigned by Audit Analytics for each auditor
auditor_state_name	Translation of state code field to geographic region
effective_internal_Controls	1 indicates that the auditor or management found the registrant's internal controls over financial reporting to be effective.
audit_Fees	Total audit fees paid during the Fees Fiscal Year Ended
non_Audit_Fees	Total non-audit fees paid during the Fees Fiscal Year Ended
total_Fees	Total auditor fees paid during the Fees Fiscal Year Ended
share_Price	Closing price of the registrant's equity on the date specified (see field Stock Price Date)
market_Cap	Market capitalization as of Stock Price Date
revenue	Corresponds to the past year's Total Revenue usually found as the first item on the Income Statement
earnings	Corresponds to the past year's Net Income and is a calculated field (Revenues – Expenses) usually found as a line item on the income statement
book_value	Corresponds to the year end Book Value and is a calculated field (Total Stockholders' Equity – Goodwill – Intangible Assets
assets	Corresponds to the year end Total Assets and is usually found as an item on the Balance Sheet
big_four_indicator	1 if the company's auditor is one of the Big Four firms

Variable	Description
five_category	1 if company's auditor is PricewaterhouseCoopers, 2 is Ernst & Young, 3 is Deloitte, 4 is KPMG LLP, 5 represent any non-Big Four firms
audit_fees_bc	Boxcox transformation of audit fees
total_fees_bc	Boxcox transformation of total fees
market_cap_bc	Boxcox transformation of market capitalization
assets_log	logarithm transformations of assets
revenue_trans	Non-zero processing and logarithm transformation for revenue
earnings_trans	Non-zero processing and logarithm transformation for earnings
big_4_factor	Transformation of properties of big_4_indicator variables
five_category_factor	Transformation of properties of five_category variables
state_region	Transformation of properties of state_region variables

We examine the characteristics of these variables and summarize them in the figures below:

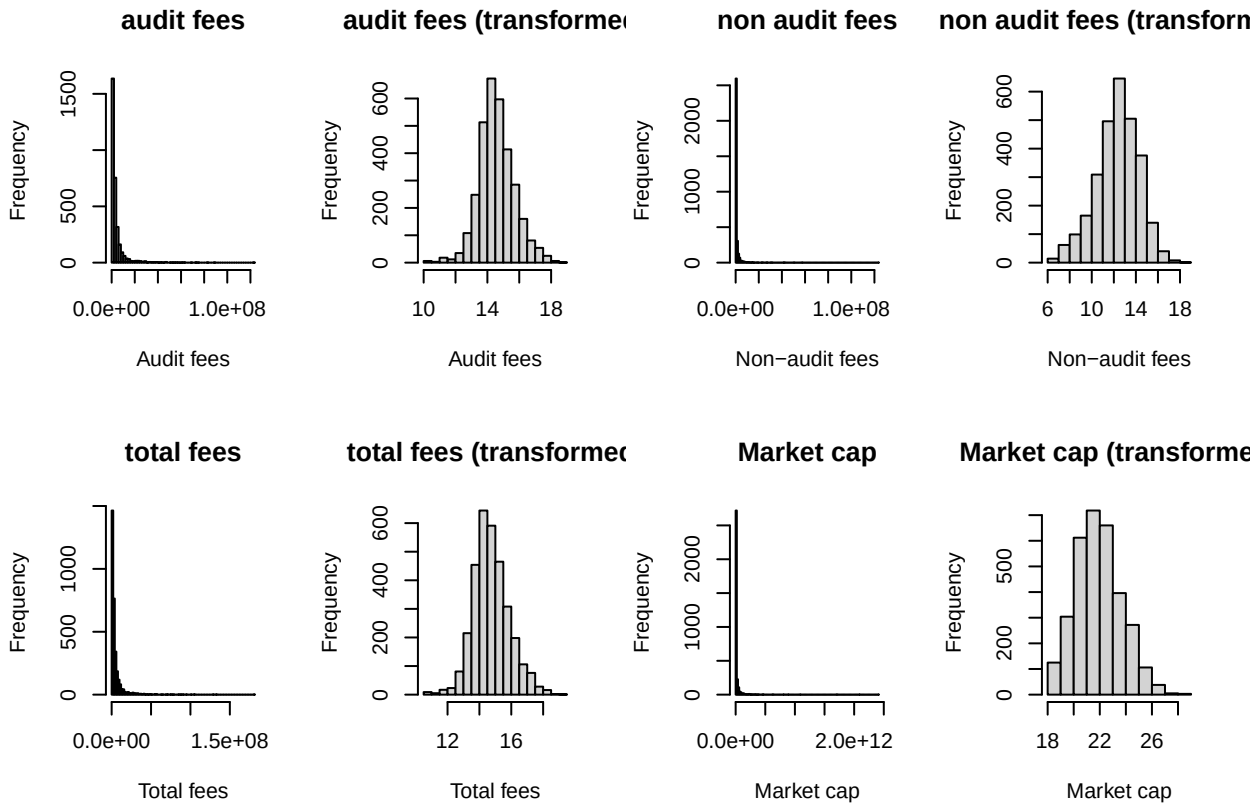

Figure 1: Variable Transformation

Figure 1 provides 4 2X2 panels about pre-transformation versus post-transformation of distribution of the public companies from the perspective of audit fees, non-audit fees, total fees as well as market capitalization. Data distribution is very skewed before its transformation, but the post-transformation distribution is more symmetric and normalized, which can help to improve the accuracy of following statistical analysis.

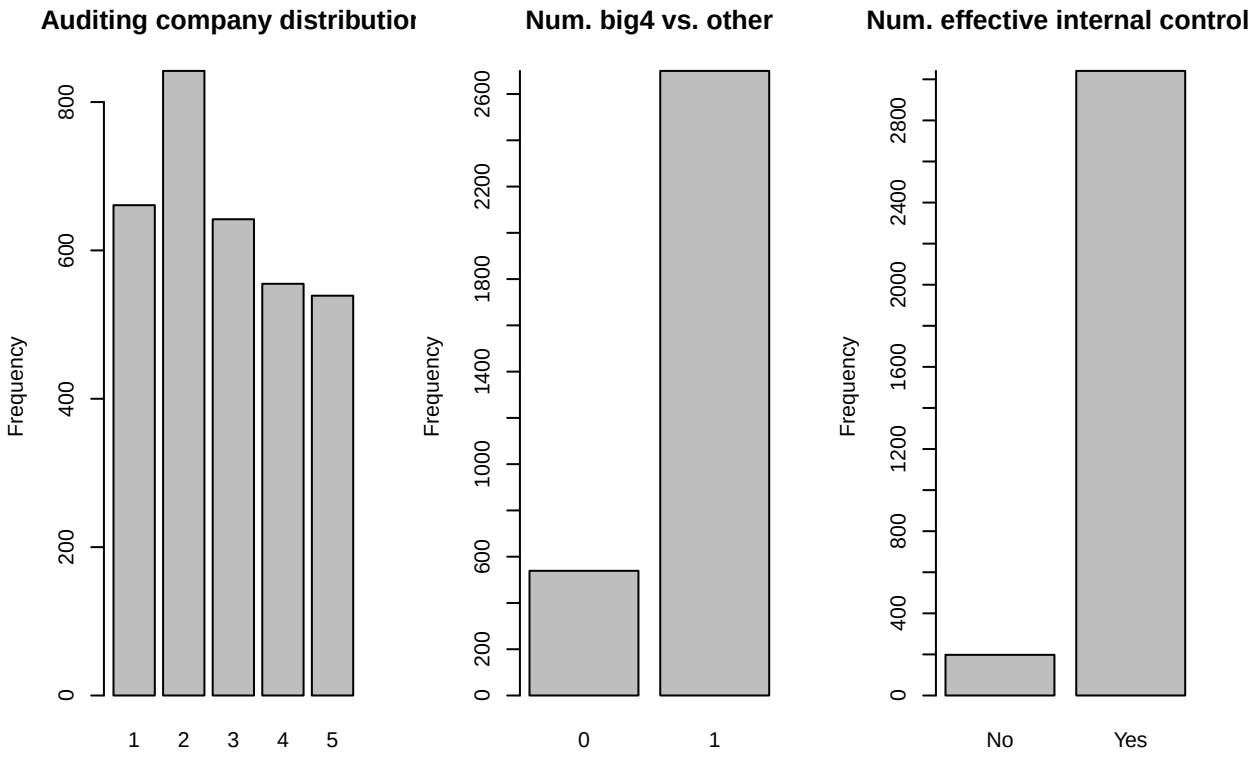

Figure 2: Distribution of Categorical Variables

The first panel in Figure 2 shows that the Big 4 firms generally have the largest market share, with each of the Big 4 firms having more audit clients than all non-Big 4 audit firms. The mid panel in Figure 3 provides further evidence for aforementioned observations. Among all companies on the U.S. stock exchanges, approximately 6.25% of them have non effective internal control(i.e. receive adverse audit opinion on ICFR) in 2021 fiscal year, which is showed in the third panel in Figure 3.

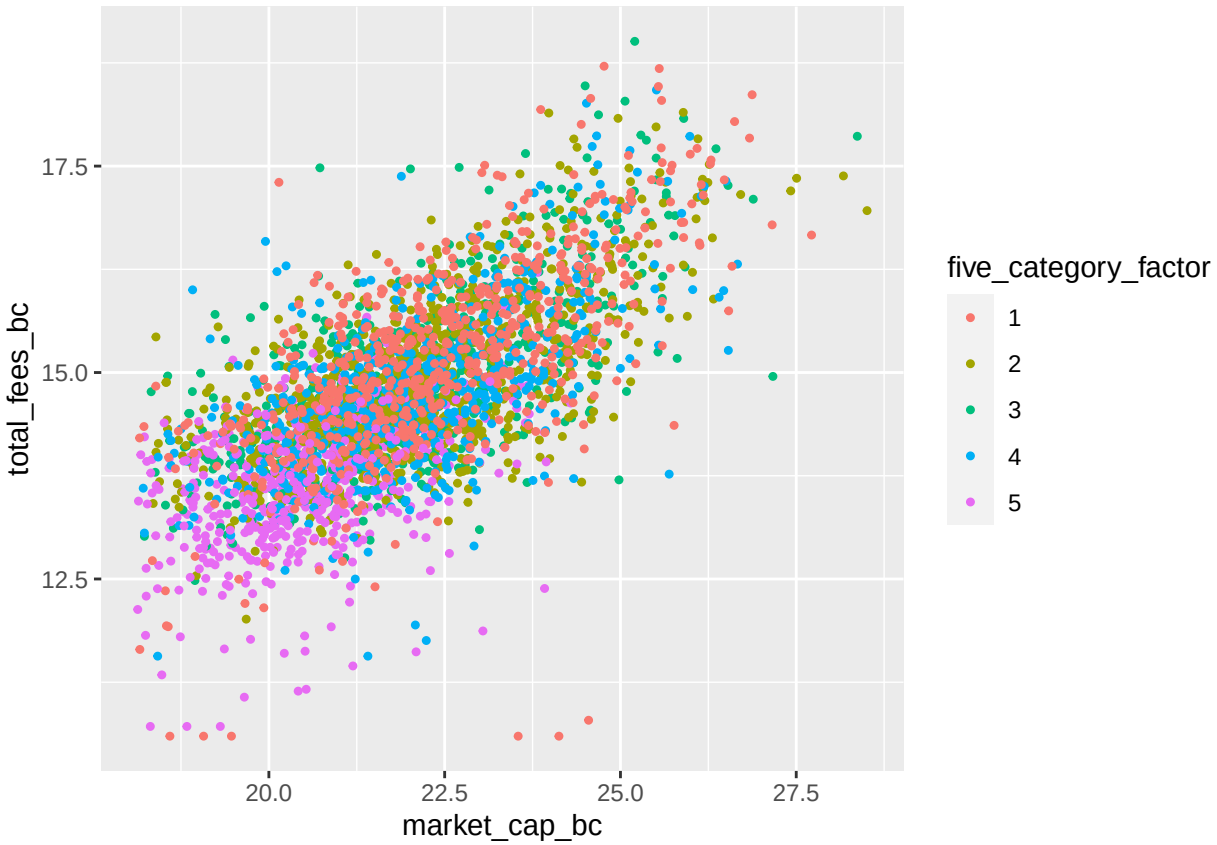

Figure 3: Relationship between Total Fee and Market Capitalization

The correlation parameter between market cap and total fee is 0.69. From the Figure 3, we can see that the non-Big 4 firms are obviously on the lower left corner. This represents that market cap of the companies which are audited by the non-Big 4 is lower, and the total fee the company pay for audit is also lower. That's reasonable because the Big 4 has their brand value.

We noticed that sometimes the original data have abnormally low audit fees (errors may be due to wrong allocation of audit fees and non-audit fees in the original data set). Since in the data set, audit fees account for a large proportion of the total fees of each public company, using total fee can represent the audit fee better. Aiming to have a more accurate analysis, we use data of total fees instead of audit fees in the following analysis, while the name of total fee is interchangeable with audit fee.

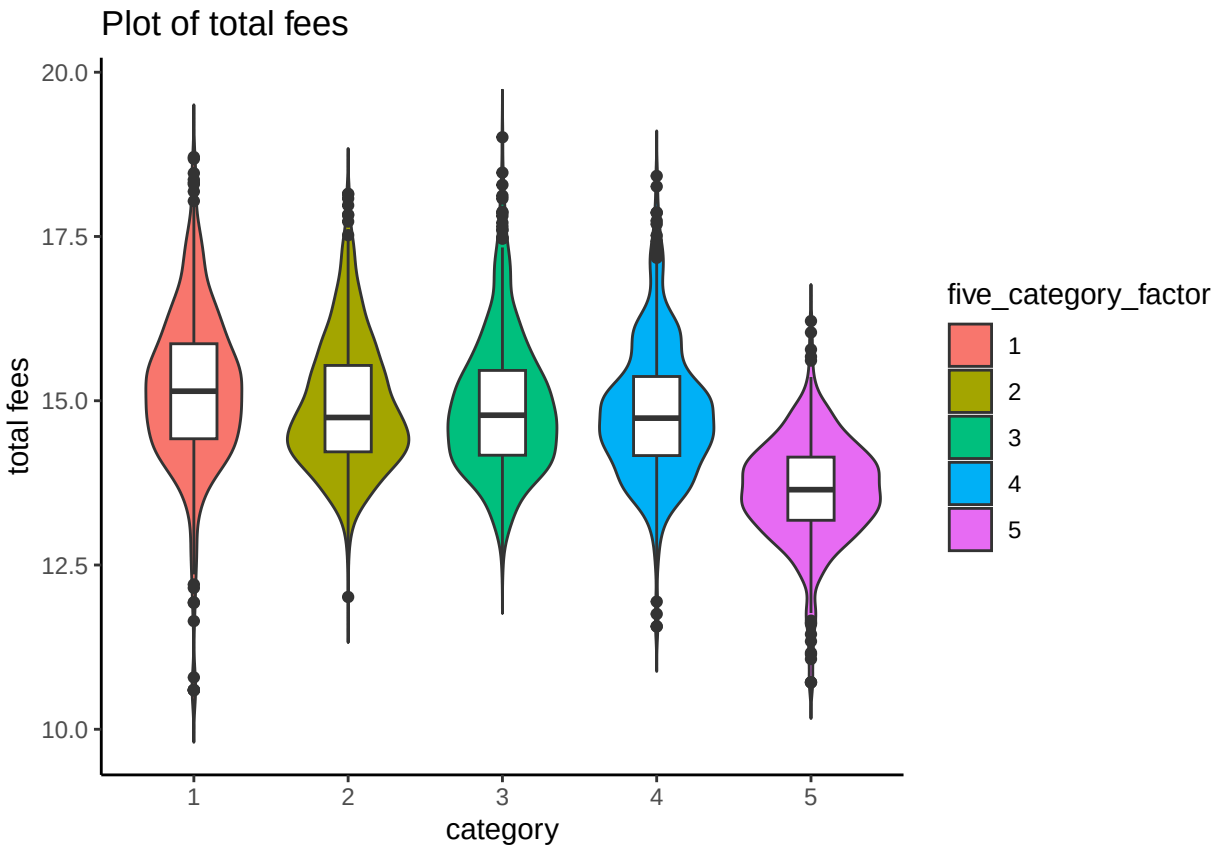

Figure 4: Total Fee Comparison

Figure 4 shows that generally all Big 4 firms have higher median total fees than non-Big 4 firms. We can see that the highest total fee is charged by one Big 4 firm, Deloitte and the lowest total fee is also charged by another Big 4 firm PricewaterhouseCoopers LLP instead of non-Big 4 firms. Whether the auditor is one of the Big 4 firms isn't the only factor determining total fees. Therefore, what are the key factors to determine total fees need predictive model to explore.

Perform a statistical test here to compare big4 vs non big4 when considering

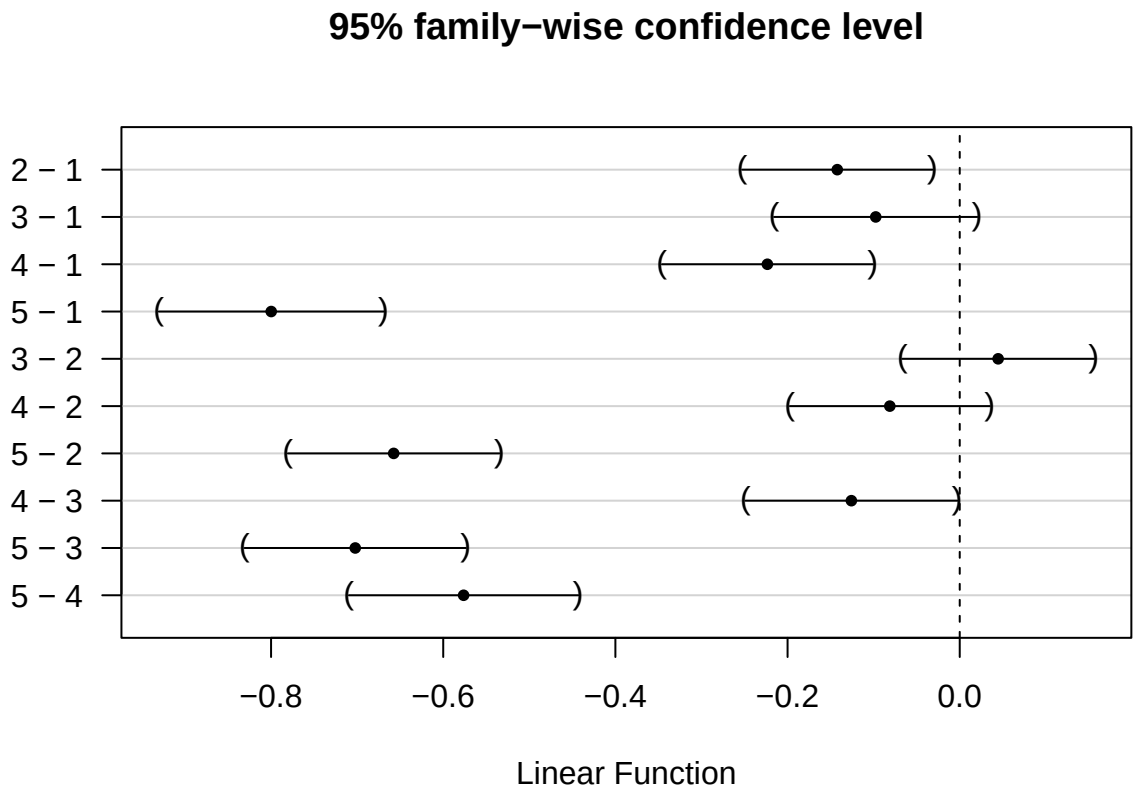

Figure 5: Average Audit Fee Comparison Between Audit Types

Figure 5 shows the Tukey's test outcome, which compares the means of audit fees of all pairwise comparisons among every Big 4 firms and non-Big 4 firms as a whole. We can see that controlling market capitalization, average audit fees difference between any pair of Big 4 firms is little while average audit fees difference between any Big 4 firms and non-Big 4 firms are statistically different.

Those above analysis arouse our great interests to explore what are the key factors to determine audit fees and we try to find our answers by the following steps.

Unsupervised Learning: Clustering

After exploring the data set, we continued to cluster with the companies based on the following values before and after previous transformation: audit fees($), total fees($), Market capitalization($), market-fee ratio, asset value($), revenue($) and earnings($). We used the Gap Statistic and K-means methods to determine the number of clusters.

Gap Statistic for Estimating the Number of Clusters

This method estimates a goodness of clustering measure, the "gap" statistic with a given range of number of clusters K. For each K, it compares the log value of dispersion of observations within a cluster to the estimated log value of dispersion. In this report, the maximum number of clusters to consider is 10.

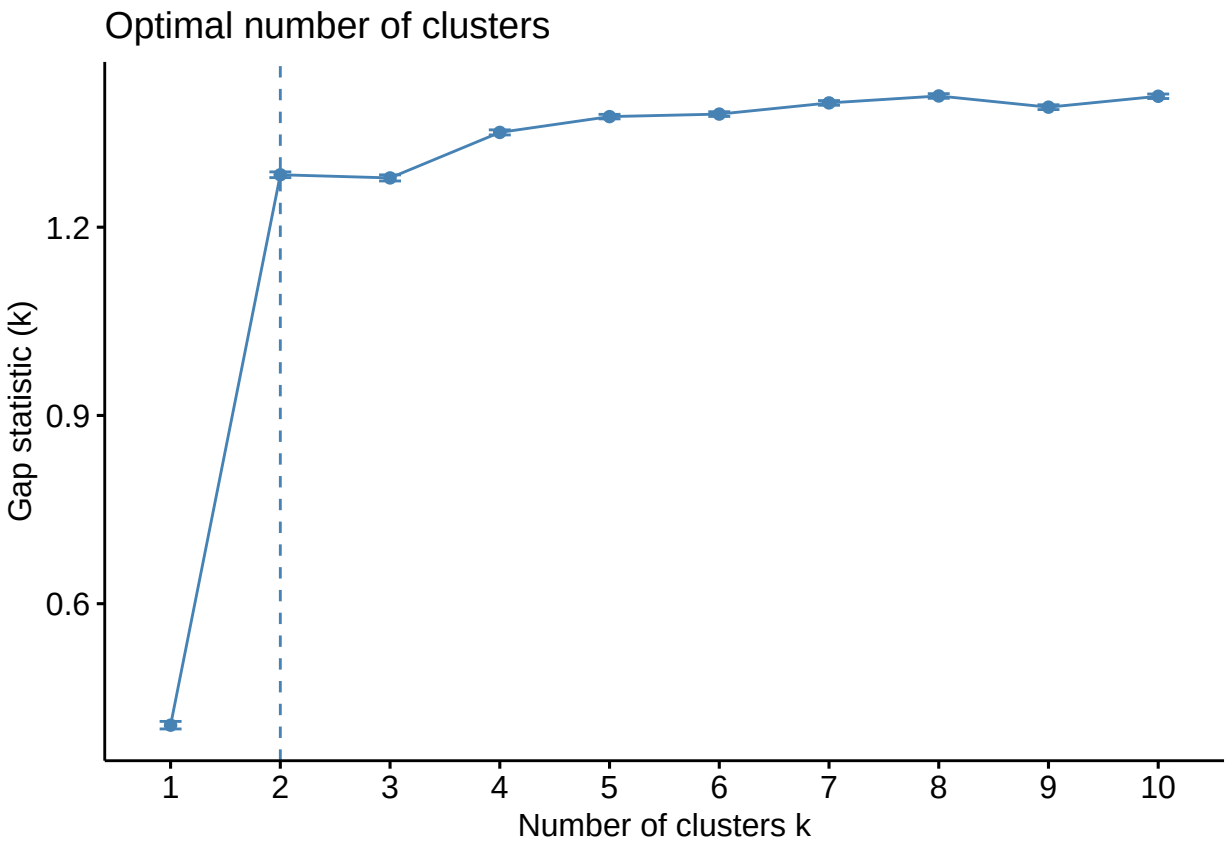

Figure 6: Gap Statistic Clustering Result

As The Figure 6 suggests, the Gap statistic estimates that the optimal number of clusters is 2.

K-means Clustering

The K-means clustering uses a principal component analysis to create clusters, and classify and partition objects into multiple groups. The objects are as similar as possible within, and as dissimilar to the objects in other groups as possible. K represents the number of groups. From the previous result, we select 2 sets of groups, with the same variables as Gap statistic used.

2 clusters are generated with distinct features. The Figure 8 shows that these two groups overlap each other in large areas, which is expected because the box-cox and natural log transformation transformed value of the data to be closer to each other. Unsatisfied with the result, we conduct the same clustering with original data.

Table 2: Transformed Features of K-means Clustering Groups

cluster	audit_fees_bc	total_fees_bc	market_cap_bc	market_fee_ratio	assets_log	revenue_trans	earnings_t
1	14.334	14.469	21.169	-6.700	21.040	17.655	-18
2	14.682	14.845	22.196	-7.352	22.399	21.212	19

Table 3: Untransformed Features of K-means Clustering Groups

cluster	audit_fees	total_fees	market_cap	market_fee_ratio	assets	revenue	earnings
1	3948548	4824005	17136157968	-7.174	2.111079e+10	7584571466	816700052
2	58787475	71092557	122335978468	-6.804	2.251131e+12	52604000000	15833555556

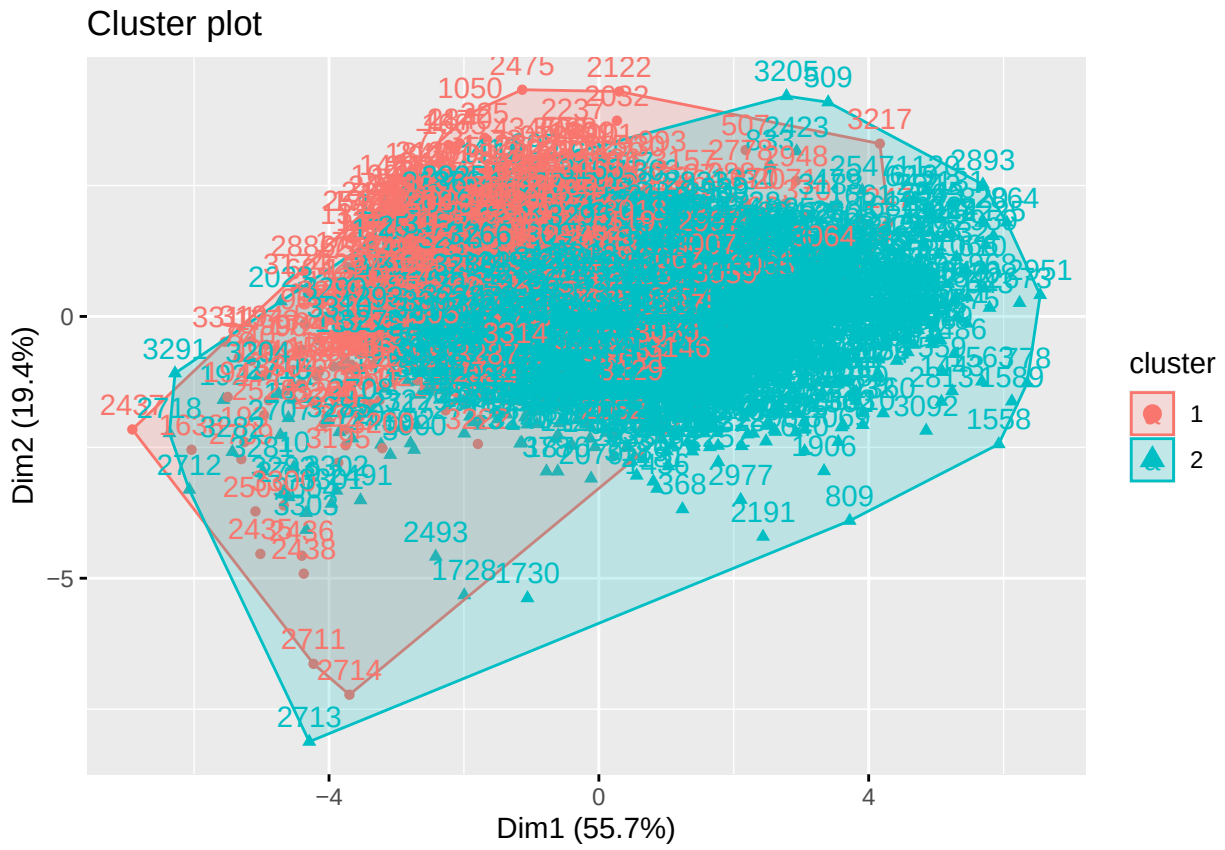

Figure 7: K-means Clustering Result of Transformed Data

As shown in Table 3, companies in one cluster have smaller market caps and assets than companies in the other cluster. In the higher cluster, the mean asset value is $2.2 trillion and mean market capitalization is 1.2 trillion. The lower cluster has mean value of asset worth $0.2 trillion and 0.17 trillion market capitalization. As a result of less market capitalization, companies in the higher cluster have a mean market-fee ratio of 0.12% while companies in the other cluster has 0.06%. This cluster also has less revenue and earnings than the higher cluster. For instance, mean revenue of the higher cluster is approximately $0.78 trillion while the mean revenue of the lower cluster is $5.2 trillion.

In Figure 7, The component of the first principal component accounts for 55.8% of the total variation, and the second principal component accounts for 19.4%. Together they explain 75.2% variation in the data set. The same component in Figure 8 explains 72.8% variation. The limited number of variables selected when clustering can explain this issue. Along with variables' values that are similar to each other after transformation, the same reason could explain the large area that clusters overlap each other in Figure 9. The overlapped area is smaller when the clusters are constructed with untransformed data and they are more distinctive.

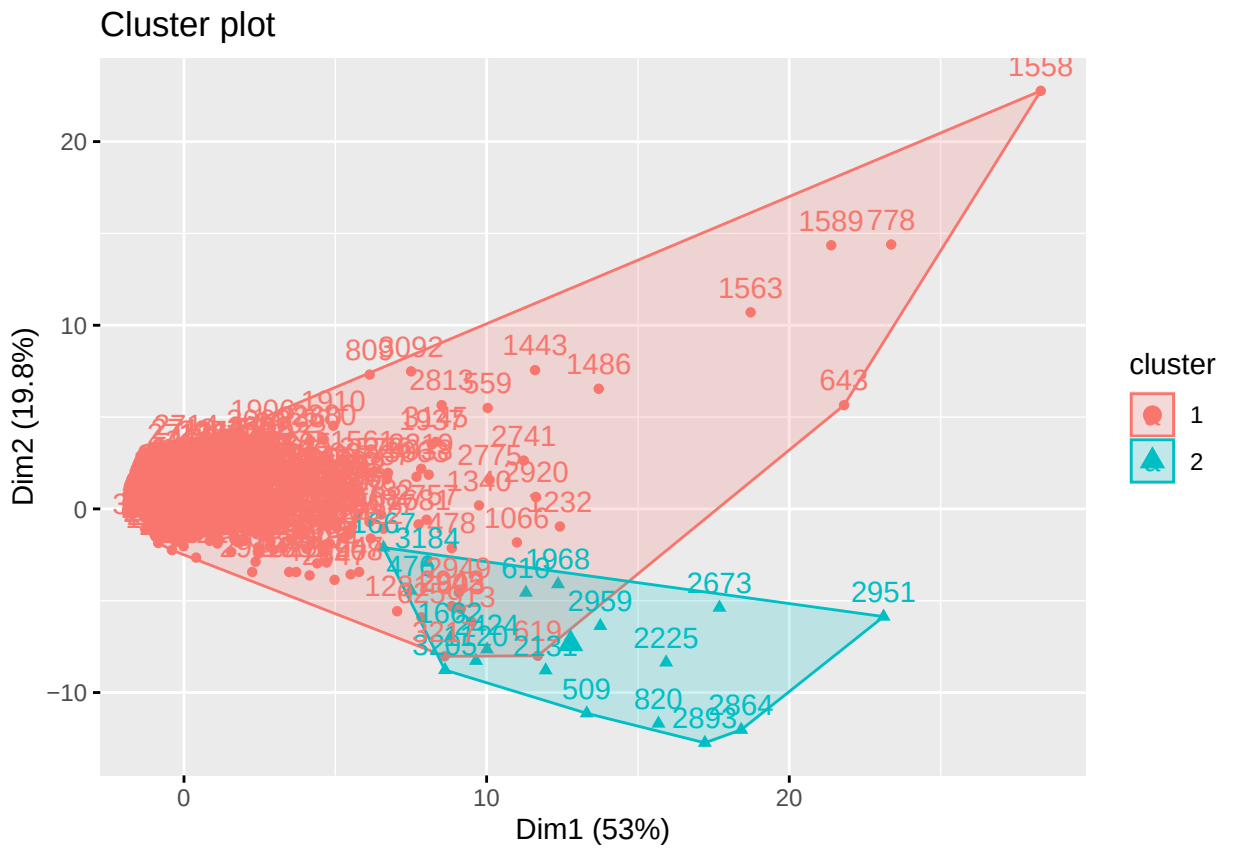

Figure 8: K-means Clustering Result of Untransformed Data

Supervised Learning

We noticed that the original data have abnormally low audit fees (errors may be due to wrong allocation of audit fees and non-audit fees in the original data set). Since audit fees account for a large proportion of the total fees of each public company, aim to have a better analysis, we will use data of total fees instead of audit fees in most of our following supervised learning models.

In the supervised learning part, we want to build several models and choose the best performance model to do the prediction of total fees. The workflow is shown below:

Firstly, we did the preliminary feature engineering, we abandon some trivial features. For example, the name of the company, the city name and the state name, etc. We want to try various models. We chose 5 models here: Linear model, KNN model,Random forest model, CART model and Gradient-Boosting model.

Secondly, we used K-fold cross validation to assess the model.

Lastly, we would consider the MAE, RMSE and R-square to decide which model has the best performance.

```
##
## Call:
## summary.resamples(object = res)
##
## Models: lm, knn, rf, cart, gbm
## Number of resamples: 5
##
## MAE
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm   0.5000683 0.5059538 0.5113339 0.5147180 0.5152495 0.5409847    0
## knn  0.4424454 0.4475611 0.4528507 0.4569064 0.4645350 0.4771399    0
## rf   0.4235247 0.4271173 0.4312494 0.4300904 0.4337744 0.4347862    0
## cart 0.5047285 0.5141932 0.5146004 0.5230124 0.5333537 0.5481863    0
## gbm  0.4205715 0.4269697 0.4353175 0.4351127 0.4415710 0.4511336    0
##
## RMSE
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm   0.6304778 0.6366461 0.6468913 0.6473935 0.6502904 0.6726619    0
## knn  0.5588373 0.5779585 0.5873867 0.5913374 0.6008683 0.6316362    0
## rf   0.5436985 0.5501216 0.5512979 0.5516267 0.5544380 0.5585773    0
## cart 0.6469405 0.6604812 0.6643343 0.6754304 0.6937969 0.7115991    0
## gbm  0.5388163 0.5562922 0.5622507 0.5598782 0.5647957 0.5772362    0
##
## Rsquared
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm   0.6564152 0.6715448 0.6764661 0.6725030 0.6783368 0.6797522    0
## knn  0.7017945 0.7057038 0.7236032 0.7262060 0.7368449 0.7630836    0
## rf   0.7497015 0.7557800 0.7602745 0.7616227 0.7678303 0.7745272    0
## cart 0.6201592 0.6248898 0.6378829 0.6419911 0.6509581 0.6760655    0
## gbm  0.7407988 0.7548725 0.7591123 0.7551768 0.7602268 0.7608734    0
```

Depending on the above table, after the advanced feature engineering, we see that random forest model has the best performance, which has the lowest MAE, lowest RMSE and highest R-square.

So we choose Random Forest model to do the prediction of total fee based on other variables.

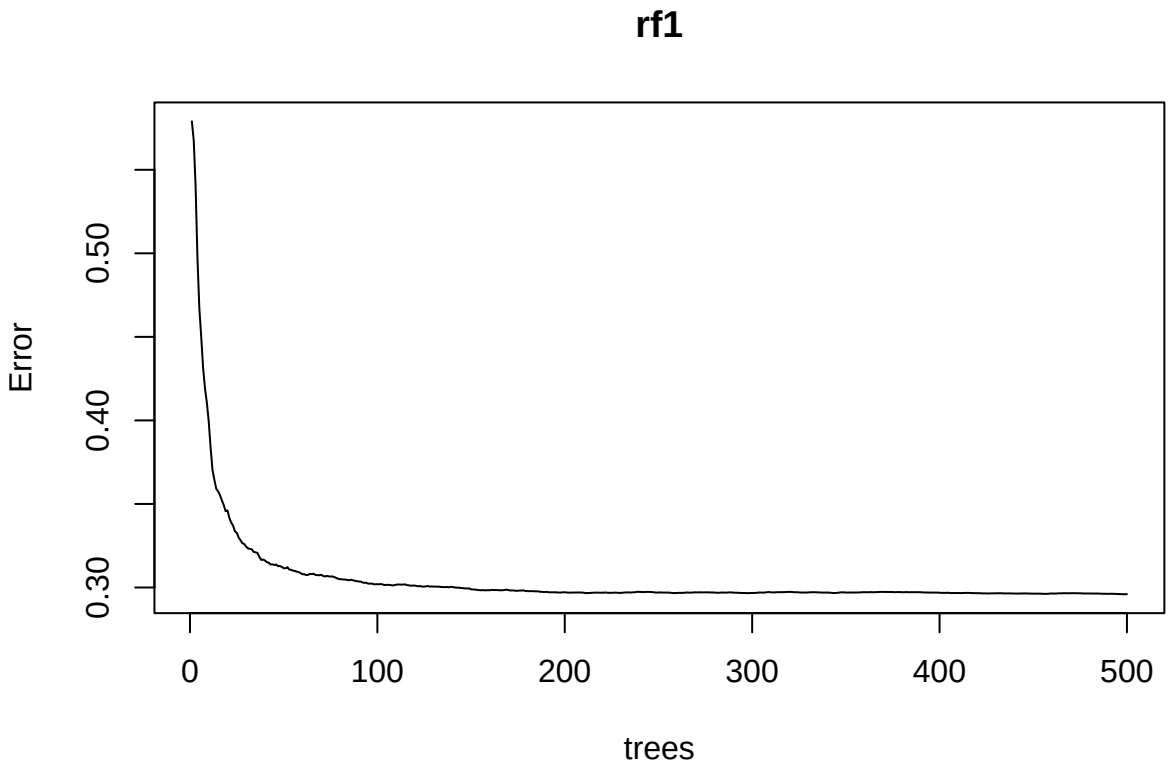

Figure 9: Relationship Between Error and Number of Trees in Random Forest Model

The plot of rf1 shows out-of-bag MSE as a function of the number of trees used. When the number of tree is equal to about 200, the MSE is slowing down. So we test number of tree = 200.

```
## RMSE of 200 Trees =  0.5858147
```

```
## RMSE of 500 Trees =  0.5858135
```

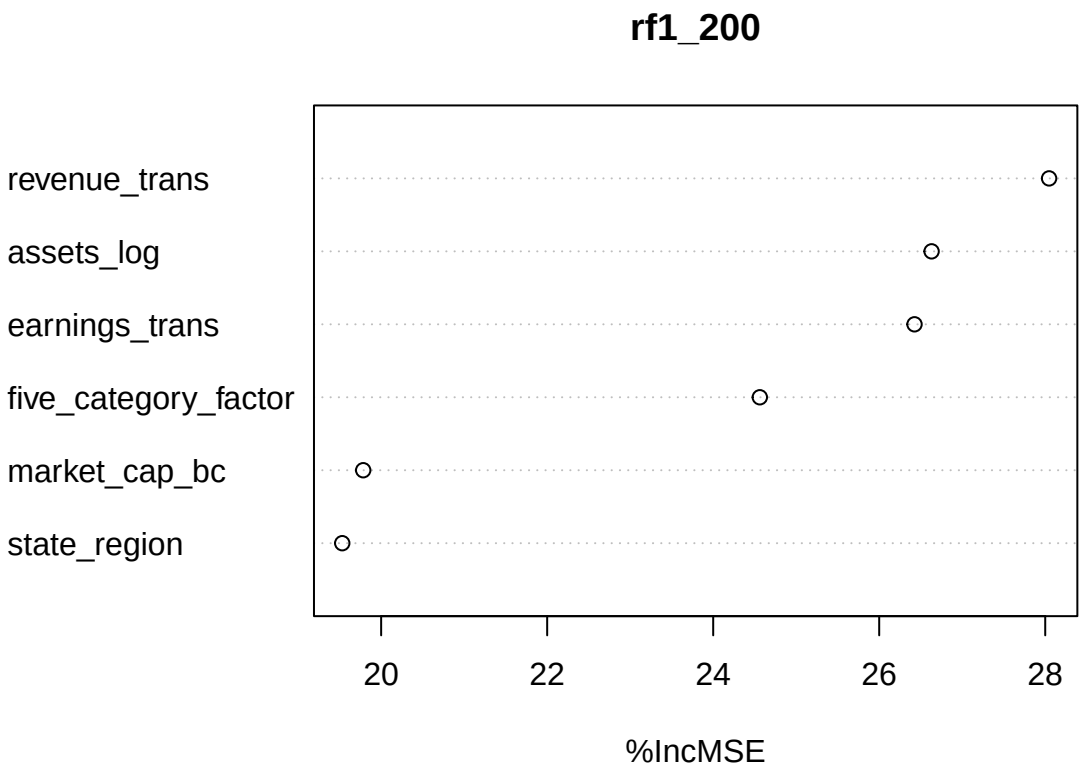

Figure 10: Influential Variables on Total Fee ranked by IncMSE

```
##   Tree      RMSE
## 1  200 0.5858147
## 2  500 0.5858135
```

Then we see the RMSE of random forest model(rf1_200) is 0.5864, which is pretty similar to that of tree=500. Choosing tree=200 could be a better choice to make the random forest model more concise. In the Introduction part, we analyze the complexity of public companies as well as expertise and resource of audit firms are important factors for audit fees. Thus we chose revenues, earnings, assets and market caps as the characteristics of the company in our above model. Additionally, we add five category factor as the indicator of the brand of the audit company: 1-4 for Big 4 firms and 5 for non-Big 4 firms in our above model. From the outcome we got, we can see that revenues_bc has the most influence on the total fees.If we delete the revenue variable, the accuracy of model will decrease about 28%. The other 5 variables: the region of the company, the market cap of the company, the asset of the company, the earnings of the company and whether the company is audited by the big 4, also have major influence on the model.

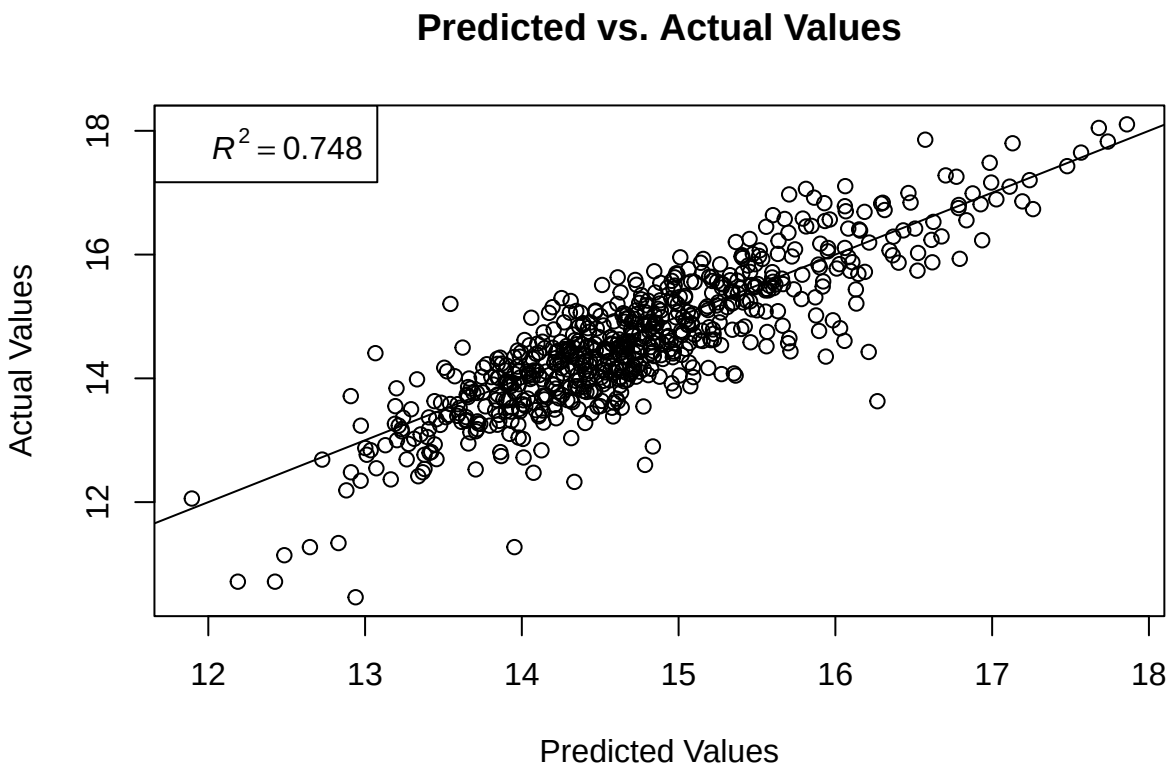

Figure 11: Performance Evaluation of Predictive Model

From the above figure of predict value v.s. actual value, we saw that our model has good performance. The R-square is 0.748, which means that about 75% of the variability observed in the total audit fee can be explained by the random forest model. However, we know that we can't just use R-square to assess the model performance. We also consider MAE and RMSE. Compare to other models, random forest model still has lowest MAE(the average absolute magnitude between the actual values and the predicted values) and RMSE(the average difference between values predicted by a model and the actual values). As a result, we can make sure that random forest model is the best performance model.

Conclusion

Companies listed on U.S. stock exchanges hire independent auditors to audit their financial statements annually. Usually, an audit fee is paid to the auditor and covers the cost of their services. In this project, we explored and determined key factors that have critical impact on audit fee decision made by auditors, as well as predicting audit fee with several parameters. We explored the internal control data of numerous companies with market capitalization above 75 million in 2021, used unsupervised learning method and found that these companies can be clustered to two groups with distinctive features. Supervised learning method allowed us to select Random Forest model that fits the best and predicts audit fee the most accurately.

Determining accurate audit fee is arguably one of the most important components in auditing process. This project contributes to predicting accurate audit fee, which gives companies a good reasonable reference during the negotiation with independent auditors, it also helps auditors to assess their competitiveness in the market.

Any predictive model is built based on the data given to learn, so is ours. Our result suggests that add more variables that could affect decision of audit fee, such as restatement history, previous auditing history and companies' audit fees across a consecutive time frame will build more accurate and interesting predictive models. The data we used to train our models is limited, including companies with market capitalization below 75 million will potentially increase the model's predictive accuracy.

Citation

1. *Audit analytics*. Audit Analytics an Ideagen solution. (n.d.). Retrieved April 23, 2023, from https://www.auditanalytics.com/0002/company-search-ic.php

Appendix:

The auditor's report on internal controls over financial reporting (ICFR) is one of the few publicly observable ways for auditors to disclose unfavorable audit findings. Companies listed on U.S. stock exchanges with market capitalizations greater than $75 million must secure an independent audit opinion on ICFR under the Section 404(b) of the Sarbanes-Oxley Act of 2002, as amended by the Dodd-Frank Act of 2010. ICFR is the processes that ensure reliable financial reporting. Thus, while the auditor's opinion on the financial statements covers the product of financial reporting, the ICFR audit covers the processes that generate financial reports. The distinction is important because nearly every U.S. public company receives an unqualified (i.e., clean) audit opinion on its financial statements as the U.S. Securities and Exchange Commission (SEC) will not allow companies to trade on a stock exchange unless they correct any material errors the auditor detects in the financial statements. However, even if the audit opinion on the financial statements is clean, companies can still get an adverse audit opinion on ICFR if the auditor concludes that the company's internal controls over financial reporting are subject to one or more material weaknesses. An adverse opinion on ICFR indicates that the auditor believes that there are significant deficiencies in the company's internal controls that could potentially lead to material misstatements in the financial statements, which means that the auditor does not have confidence in the reliability of the company's financial reporting processes. Therefore, when auditors issue an adverse opinion on ICFR, auditors need to spend additional time, effort, and expertise required to mitigate the risk which could lead to material financial misstatements.