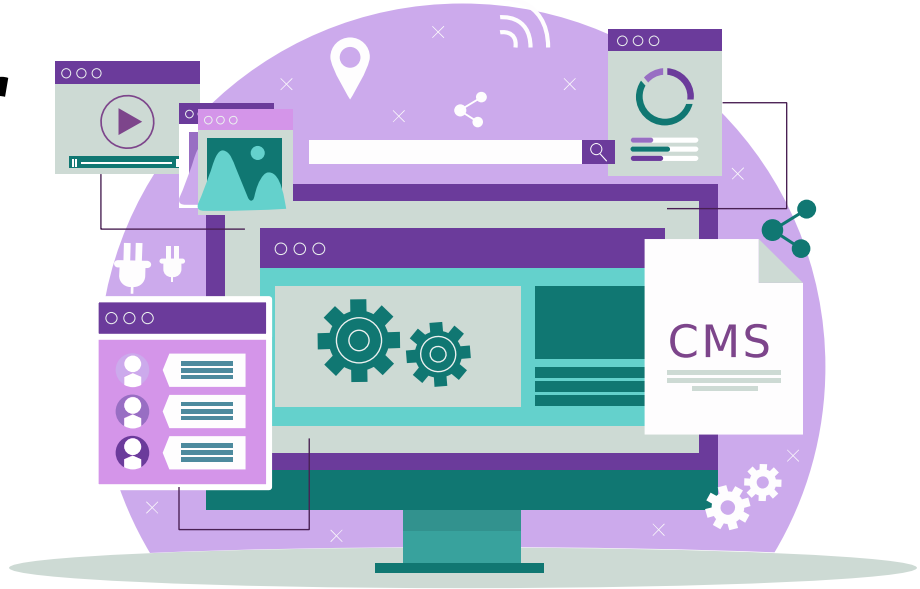
Audit Fee-ver

Team 6
Chan Hoi Yan
Chua Xin Ni
Goh Sze Pei, Cheryl
Heng Wei Shin
Ong Ken Jin



Problem Statement:

Provide a suggestion to auditors on how to set audit fees properly.

Introduction

Before we dive into our data analysis, let us learn more about what Audit Fees are.

Brief Context: What are Audit Fees?

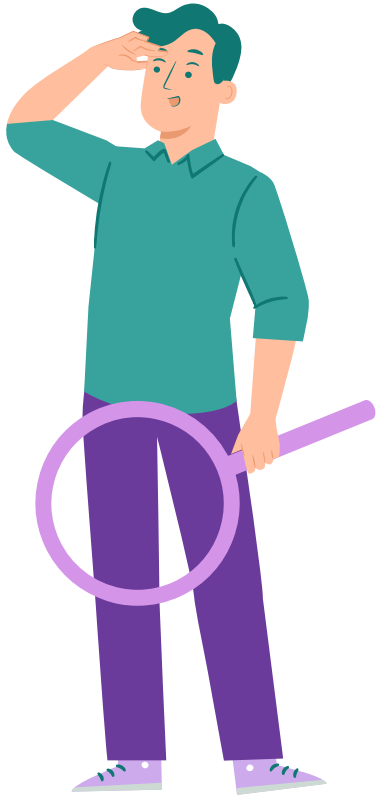
Audit Fees are costs incurred by companies to pay public accounting firms to audit the company's financial statements.

Importance

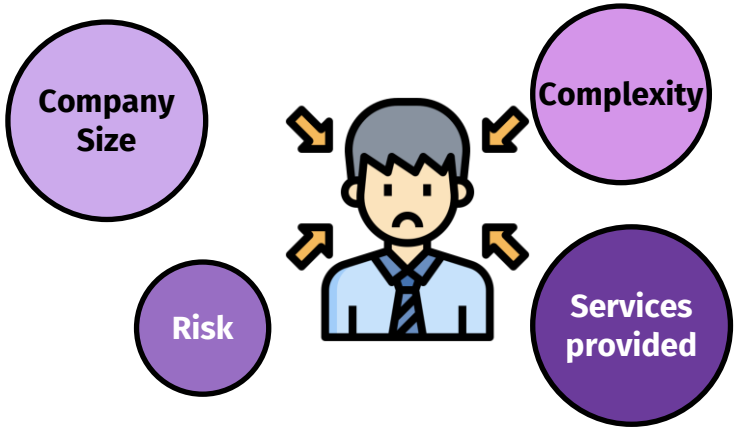
May influence an auditor's independence

Eg. Auditors may be pressured to reduce inappropriately the extent of work performed to reduce fees

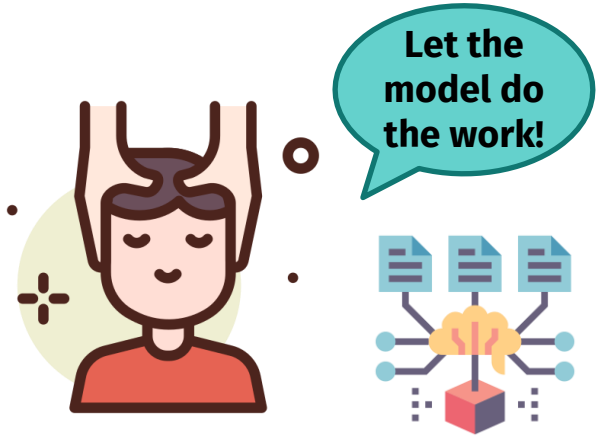
To safeguard against threats to auditor's independence, disclosure of nature of services provided and extent of fees charged is advised (though not mandatory)



Time-consuming to manually calculate audit fees



Without our model



With our model

Content Page

01

Data Explore & Clean

Overview on how we pre-processed the dataset

02

Data & Text analysis

Exploratory Data Analysis

03

Regression analysis

Key variables identified that are correlated to Audit Fees

04

Classification Tree analysis

Key trends identified by the decision tree

05

Summary

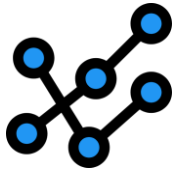
Summing up all our analysis for a recommendation

Data Preprocessing

Overview on how we pre-processed the dataset

Understanding Raw data

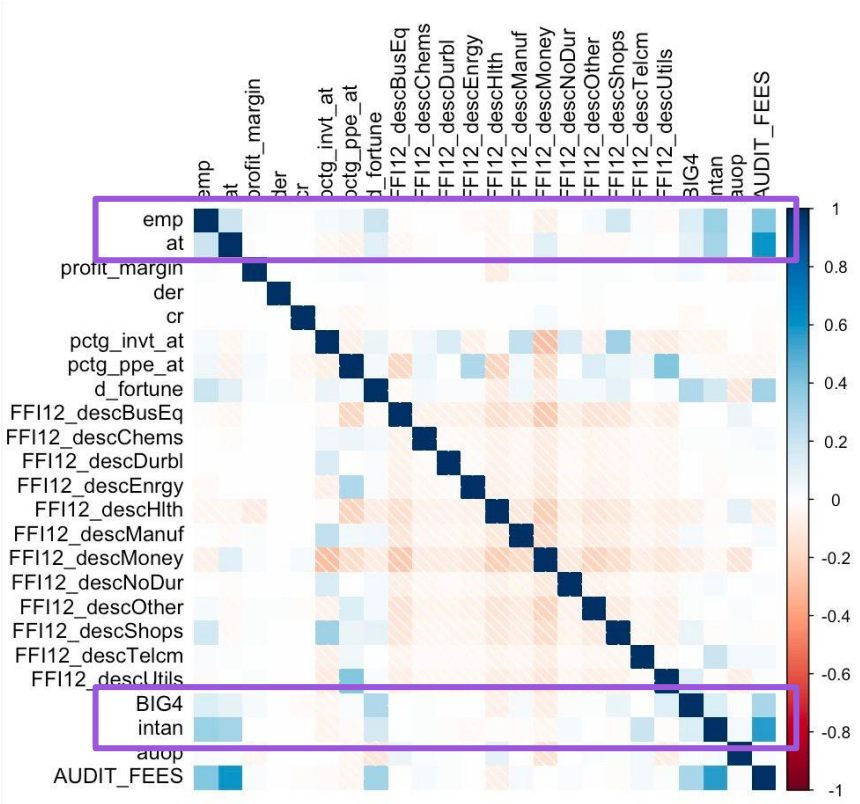
38,200 observations
5504 companies
12 industries
7 years



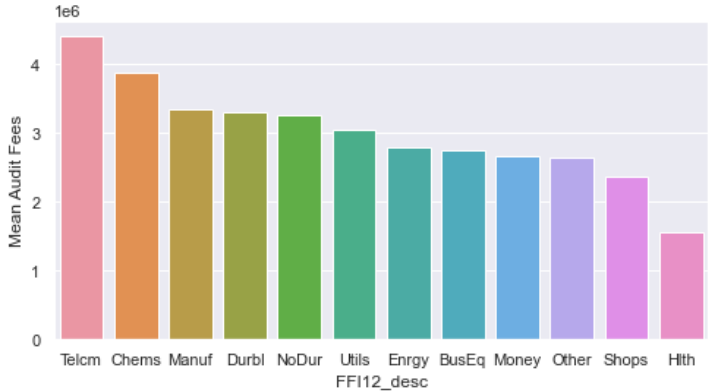
51 variables

Character Variable	FFI12_desc, busdesc
Numerical Variable	cid, datadate, fyear, fyr, act, at, capx, ceq, che, cogs, dlc, dltt, dp, dvc, emp, gdwl, ib, intan, intano, invt, lct, lt, ni, oancf, oiadp, ppegt, ppent, re, rect, sale, sstk, txdb

Taking a closer look at our data



- Audit fees are closely correlated with emp, at, BIG4, intan
- Not correlated with the types of industry, auop



Telcm has the highest Mean Audit Fees

Data Cleaning

Disclaimer: We cleaned the entire dataset but for the sake of brevity, we will only show the cleaning methods for the variables we used.

1) Replacing with company average

Before

CID	sale	total liabilities	inventory
001	NA	900	450
001	300	560	NA
001	910	NA	140
002	840	200	110

After

CID	sale	total liabilities	inventory
001	605	900	450
001	300	560	295
001	910	730	140
002	840	200	110

Assumption

- Condition of a company is roughly about the same each year
- Some things just don't change over time
- eg. similar capital structures over time

Variables affected

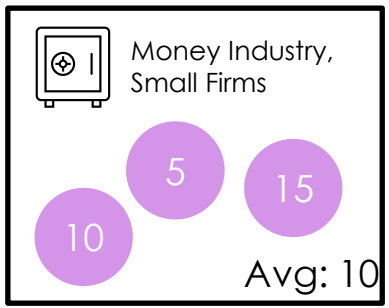
- Employee count (Emp)
- Net income (ni)
- Sales value (sale)
- Current Liabilities (lct)
- Inventories (invt)
- Short Term debt (dlc)
- Long-term debt (dltt)
- Retained Earnings (re)

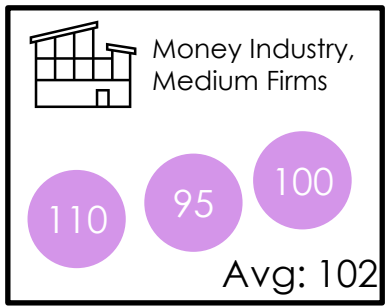
Data Cleaning

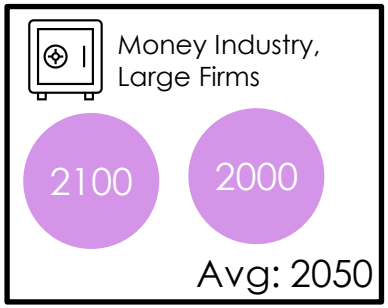
Disclaimer: We cleaned the entire dataset but for the sake of brevity, we will only show the cleaning methods for the variables we used.

2) Replacing with industry average of similar sized companies

Eg.







Assumption

- Replacing with pure industry average would disregard size of data
- Hence, used variables such as at and sales as proxies to group similar sized companies in the same industry together

Variables affected

- Employee count (Emp)
- Inventories (Invt)
- Gross PPE (Ppegt)
- Short Term debt (Dlc)
- Retained Earnings (re)

Data Cleaning

Disclaimer: We cleaned the entire dataset but for the sake of brevity, we will only show the cleaning methods for the variables we used.

3) Calculate with Industry Avg Ratio

$$lct_{cid=1} = \frac{lct}{lt}_{Industry\ Avg} \times lt_{cid=1}$$

The NA value that we want to calculate

The industry average ratio

Available data of the same company

Assumption

- Since some rules of the financial statement, like current assets cannot be more than total assets, cannot be violated, we used a relevant industry avg ratio to calculate the NAs

Variables affected

- Current liabilities (lct) – used $\frac{lct}{lt}$
- Retained earnings (re) – used $\frac{re}{ni}$
- Current assets (act) – used $\frac{act}{at}$

Data Cleaning

Disclaimer: We cleaned the entire dataset but for the sake of brevity, we will only show the cleaning methods for the variables we used.

4) Completing the Accounting Equation

Eg. Since

Assets = Equity + Liability

Also used this method to calculate new variables!
Td = dltt+dlc
Te = ceq+re

Common Ordinary Equity = Total Assets – Total Liabilities

Assumption	Variables affected
• The variables in the dataset satisfies all the logical rules of financial statements	• Common Ordinary Equity (ceq) • [NEW] Total debt (td) • [NEW] Total Equity (te)

Created Dummy variables for FFI12_desc

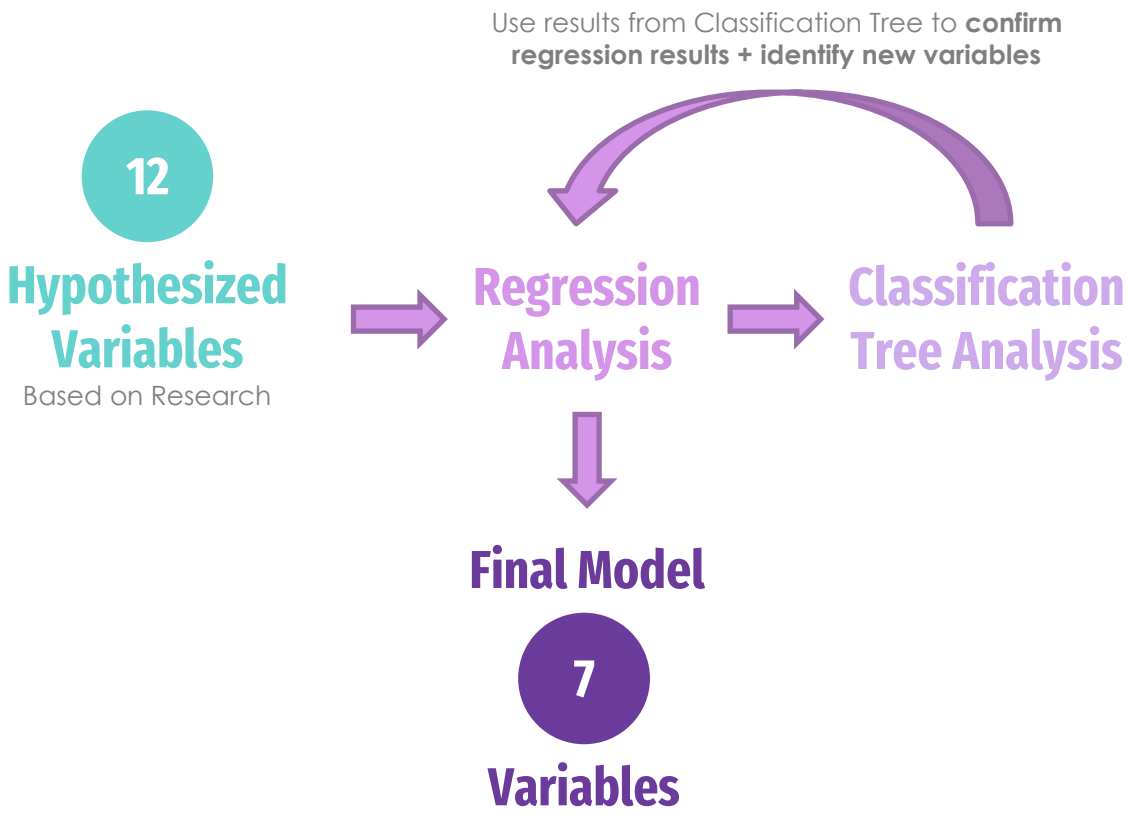
- Converted variable [ffi12_desc] into dummy variable for each industry.
- 12 additional variables are added

FFI12_descBusEq	FFI12_descChems	FFI12_descDurbl	FFI12_descEnrgy	FFI12_descHlth	FFI12_descManuf	FFI12_descMoney	FFI12_descNoDur	FFI12_descOther	FFI12_descShops	FFI12_descTelcm	FFI12_descUtils
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
											0

Explanation

- FFI12_desc given as a character variable
- However, the dataset cannot be regressed with character variables, so we converted them to numeric form

Methodology Overview



Linear Regression

Identify Key Variables that Correlate to Audit Fees

With the myriad of tasks involved in an Audit Engagement, our group has hypothesized 6 key factors that affect Audit Fees

Source(s): Academy of Accounting and Financial Studies

Breaking down how each identified factor and its proxies affects Audit Fees, we further included 2 additional variables...

Factors	Rationale
1 Status of Audit Firm	Big 4 firms charge higher audit fees due to their reputation and expertise in conducting higher quality audits
2 Industry Type	Due to the differential in complexities between industries, simple industries such as retail can expect to have lower fees than complicated industries (i.e. manufacturing)
3 Client's Size	The larger the client's size, the higher the audit fees due to the large amounts of financial data that has to be audited
4 Client's Complexity	More complex firms offering a wide variety of products and services will be harder to audit and hence charged higher fees
5 Client's Profitability	Clients with track record of low profitability has higher probability to manage earnings, more complex audit procedures required to detect risk, incurring higher fees
6 Client's Risk	Clients with higher inherent risk require larger extent of audit procedures to reduce detection risk, incurring higher fees.

We further identified two variables:

1) Amt. of Intangibles
2) Auditor Opinion

Valuation for intangible impairment assessment adds complexity for audit disclosures.

We further assume that a poor auditor opinion would positively correlate to a firm's level of risk.

Source(s): Academy of Accounting and Financial Studies, Cogent Economics & Finance

...evaluation of the proxy variables led to 6 removals due to statistically insignificant results and low R^2 value

Identified Variables

No. of Employees	Fortune 1000	Intangible Assets	Big 4 Status	Inventory % of Total Assets	PPE % of Total Assets
Total Assets	Industry	Debt-to-Equity	Current Ratio	Profit Margin	Auditor Opinion

Chosen Variables

No. of Employees	Fortune 1000
Total Assets	Industry
Intangible Assets	Big 4 Status

To be considered for final model

Rejected Variables

Debt-to-Equity	Current Ratio

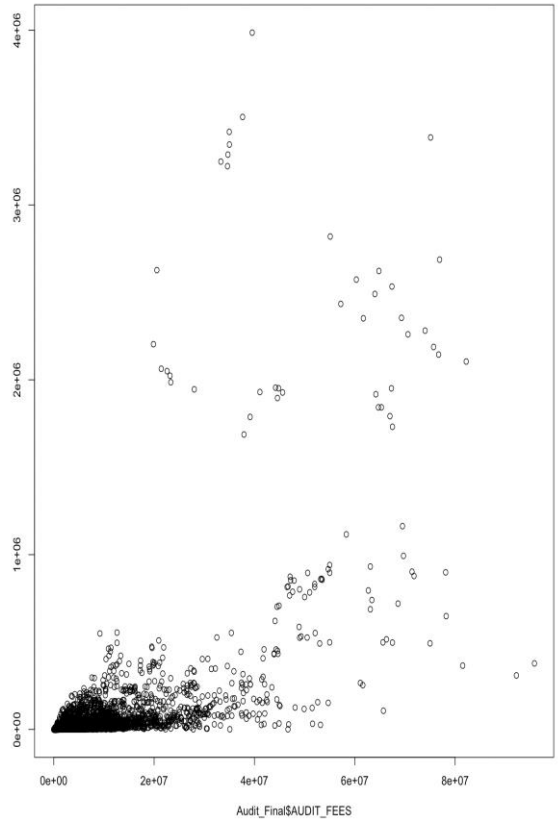
Since p-value for Debt-to-Equity and Current Ratio are 0.552 and 0.086 respectively, they are statistically insignificant at 95% CI

Inventory % of Total Assets	PPE % of Total Assets
Profit Margin	Auditor Opinion

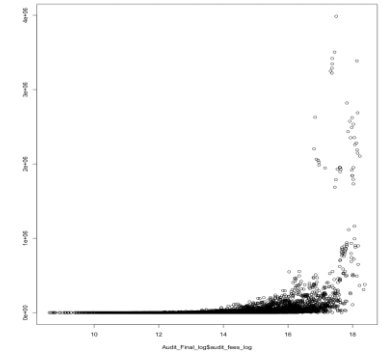
R^2 value < 1%

A strong right skew was observed for the 3 variables, resulting in the use of a log function to satisfy the linearity assumption...

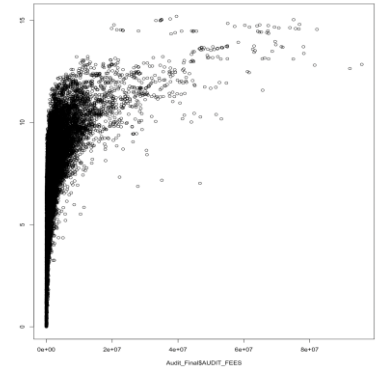
At vs AUDIT FEES



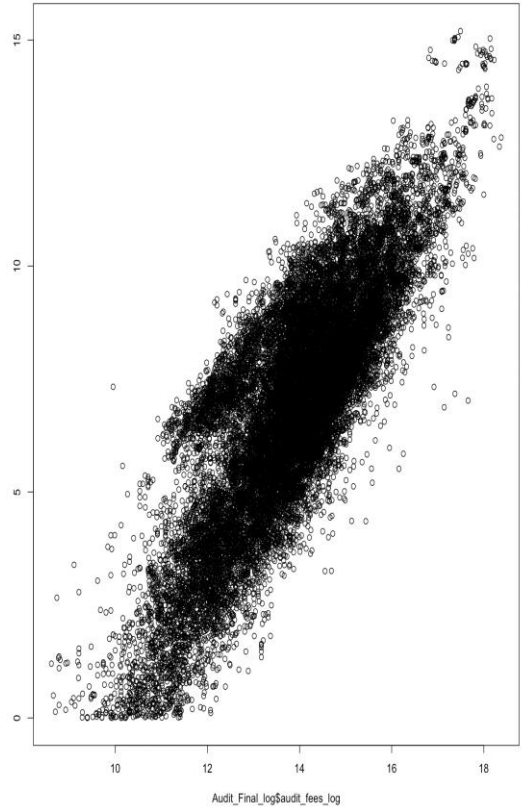
At vs log(AUDIT FEES)



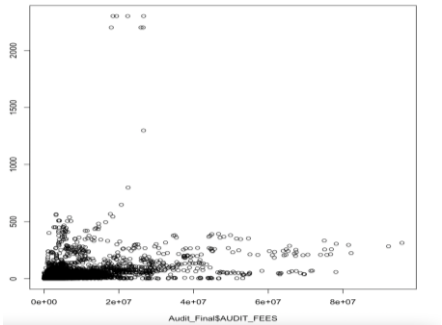
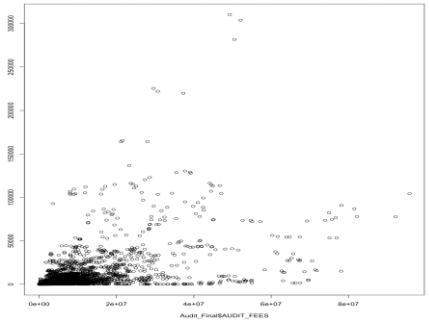
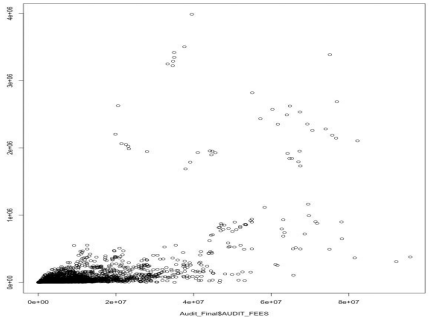
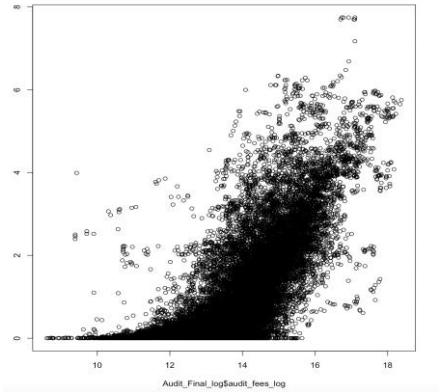
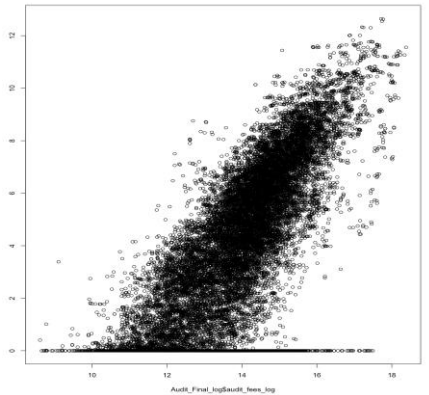
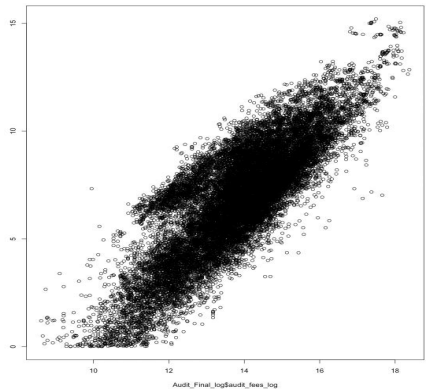
Log(At) vs AUDIT FEES



Log(At) vs log(AUDIT FEES)



...using a log function for the 3 variables results in better linearity between the variables and the log function of Audit Fees

No. of Employees	Intangibles	Total Assets	
			Without Log function
			With Log function

Minute differences between the hypothesized regression model against the variable selection models

Hypothesized Model

```
Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16
```

Forward Selection

```
Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16
```

Backward Elimination

```
Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16
```

Stepwise Regression

```
Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16
```

Across the methods, we observed that no variables were removed and hence, resulting in negligible differences between the variable selection models.

Therefore, we conclude that the model is a good predictor even without the machine learning approach.

```
Audit_RegModel <- lm(audit_fees_log ~ emp_log + at_log+ d_fortune+
                     FFI12_descBusEq+ FFI12_descChems+
                     FFI12_descDurbl+ FFI12_descEnrgy+
                     FFI12_descHlth+ FFI12_descManuf+
                     FFI12_descMoney+ FFI12_descNoDur+
                     FFI12_descOther+ FFI12_descShops+
                     FFI12_descTelcm+ FFI12_descUtils+
                     BIG4+ intan_log, data = Audit_train)
```

Classification Tree

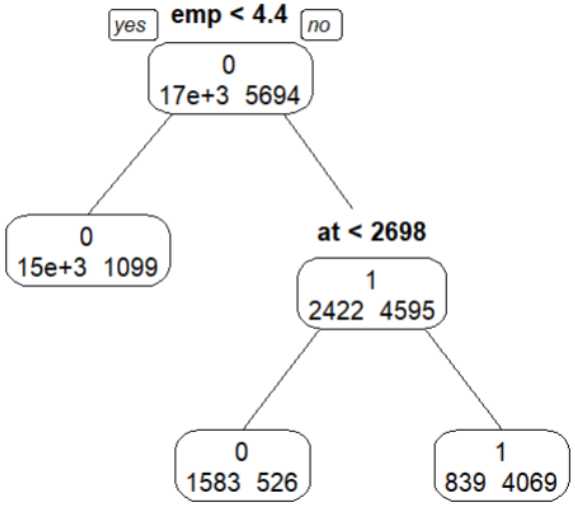
Supplement Regression &
Identify New Variables

Classification Tree: Methodology

The primary objective of the classification tree is to **cross check** against the identified variables from regression & **identify new variables**, if any.

Steps	Rationale
1 **Converted continuous variable [AUDIT_FEES] into a discrete variable via binary variable [high_AUDIT_FEES], where 1 = audit fees are above mean, 0 = audit fees are below mean**	Classification tree requires the dependent variable to be a discrete variable
2 **Ran the classification tree**	

Classification Tree 1: Mean



Results

Variables identified	Accuracy statistics
Emp at	○ Accuracy: 0.8919 ○ Sensitivity: 0.7146 ○ Specificity : 0.9509 ○ Balanced Accuracy : 0.8328 ○ Area under the curve: 0.8635

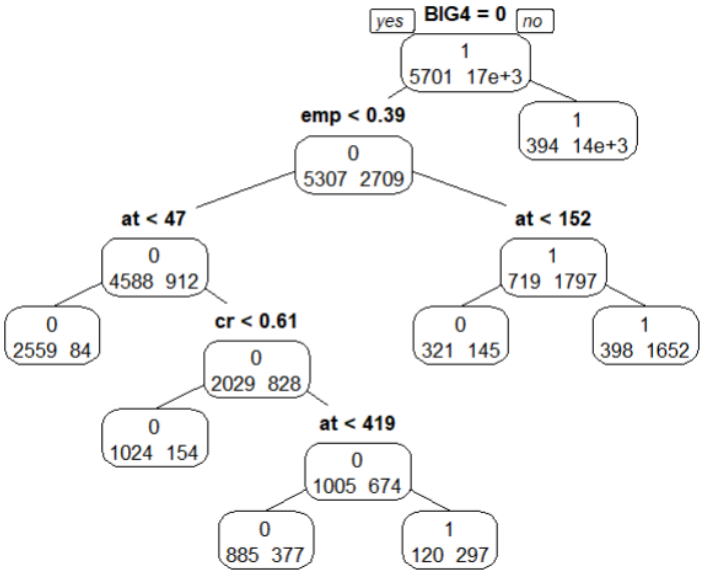
Classification Tree 1 confirmed that emp and at should be included in the final model.

Classification Tree: Further Analysis

Further analysis:
Adjusted the thresholds for **[high_AUDIT_FEES]** and ran the classification tree using these different thresholds

1. **Mean (classification tree 1)**
2. **25th percentile**
3. **50th percentile (Median)**
4. **75th percentile**

Classification Tree 2: 25th percentile

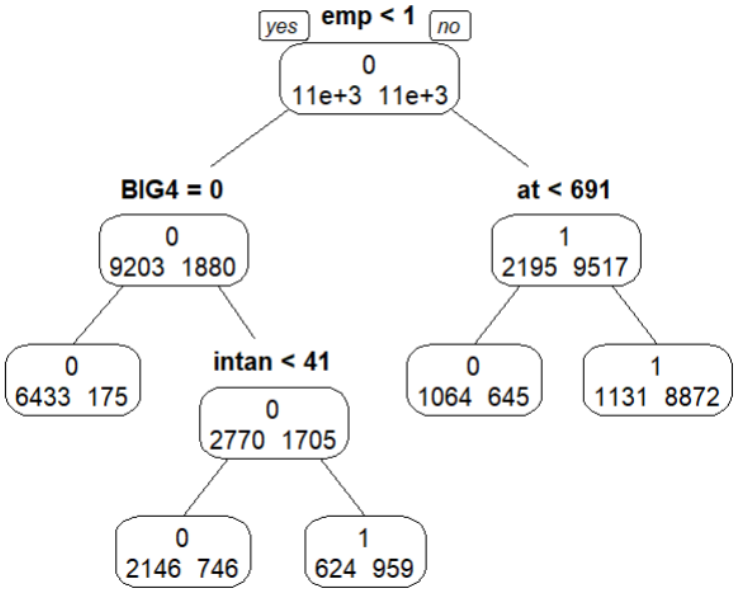


Results

Variables identified	Accuracy statistics
BIG4 (new) **Cr (new)** **emp** **at**	○ Accuracy: 0.9267 ○ Sensitivity: 0.9555 ○ Specificity : 0.8400 ○ Balanced Accuracy : 0.8978 ○ Area under the curve: 0.9422

Classification Tree 2 confirmed that BIG4 should be included in the final model. Since CR was rejected earlier due to its high p value, it will not be included in the final model.

Classification Tree 3: 50th percentile (median)

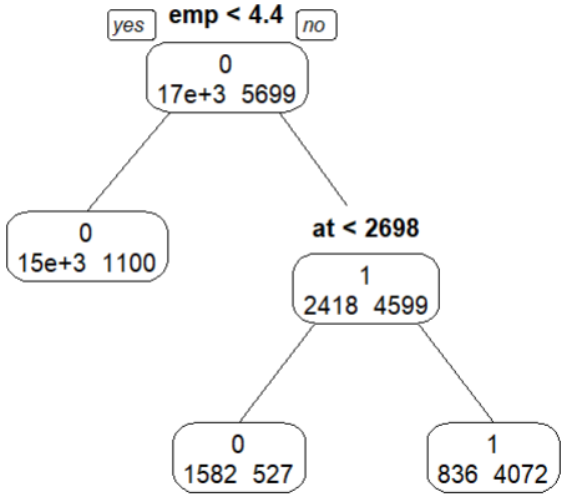


Results

Variables identified	Accuracy statistics
Intan (new) **emp** **At** **BIG4**	○ Accuracy: 0.8543 ○ Sensitivity: 0.8626 ○ Specificity : 0.8460 ○ Balanced Accuracy : 0.8543 ○ Area under the curve: 0.906

Classification Tree 3 confirmed that intan should be included in the final model.

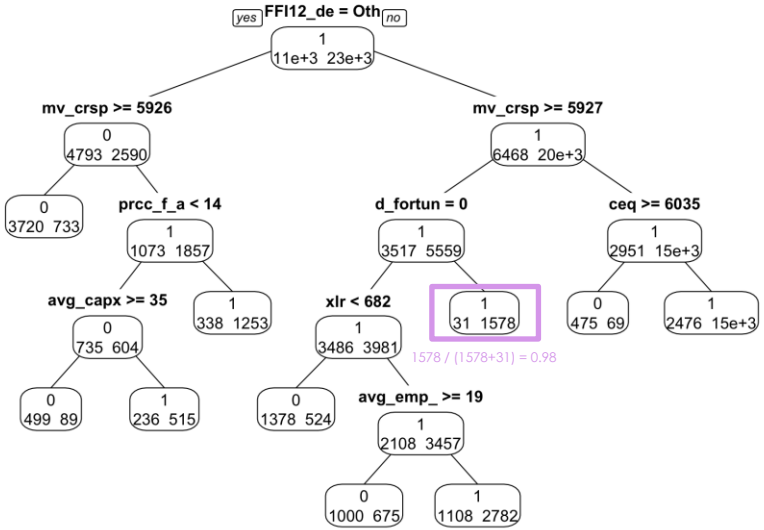
Classification Tree 4: 75th percentile



Classification Tree 4 did not identify any new variables.

Classification Tree 5: Presence of Audit Fees

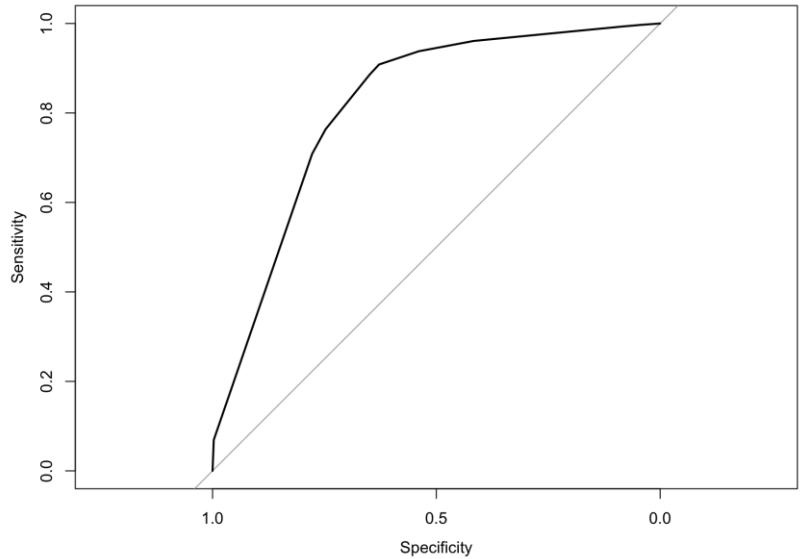
cp = 0.01824 for 7 levels of split



Methodology	Variables Identified
Dummy Variable "audited" was created *mutate(data_cleaned, audited = ifelse(!is.na(AUDIT_FEES), 1, 0))*	FFI12 != Others d_fortune **mvcrsp** (new)

Classification Tree 5 shortlisted **mv_crsp** as additional factors affecting the *presence* of audit fees…

Classification Tree 5: Presence of Audit Fees



Accuracy statistics

- Accuracy: 0.8156
- Sensitivity: 0.9083
- Specificity : 0.6280
- Balanced Accuracy : 0.7682
- Area under the curve: 0.8148

We further hypothesise that they are related to *magnitude* of Audit Fees too.

Summary

Round Up Findings &
Propose a Model for Businesses

Final Regression Model

Total Assets

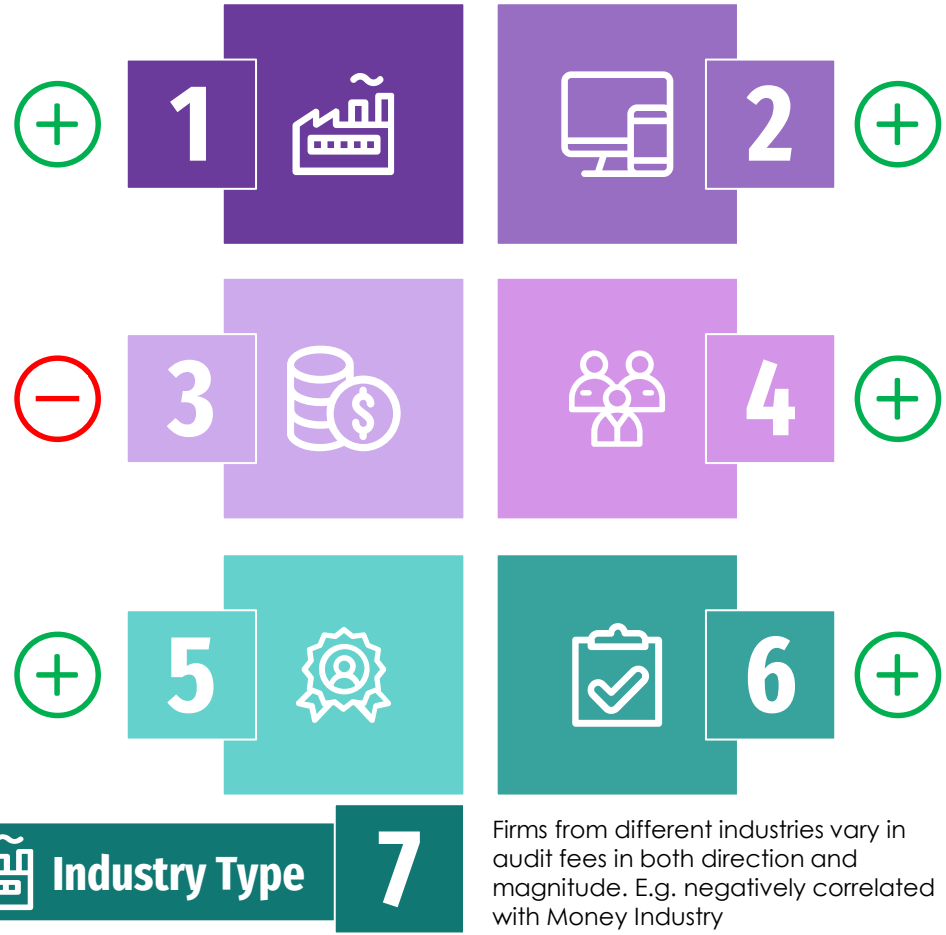
Higher Total Assets implies a larger firm with larger audit scope, increasing audit fees.
Coefficient: 0.339378

Market Value of Shareholders' Equity

Lower Shareholders Equity puts more pressure on management to perform, increasing Inherent Risk. Need to increase extent of Audit Procedures to reduce Detection Risk.
Coefficient: -0.043591

Fortune 1000?

Companies listed on the Fortune 1000 list tend to be large and subject to greater scrutiny. Auditors' scope is wider, hence increasing audit fees.
Coefficient: 0.083755



Firms from different industries vary in audit fees in both direction and magnitude. E.g. negatively correlated with Money Industry

Intangible Assets

Higher Intangible Assets may be related to higher complexity of business operations. This implies a more complex audit, thereby increasing audit fees.
Coefficient: 0.063934

No. of Employees

A higher number of employees is related to larger companies, hence larger audit scope.
Coefficient: 0.173545

Big 4 Audit Firm?

Big 4 firms tend to charge a higher fee compared to smaller, boutique firms.
Coefficient: 0.750331

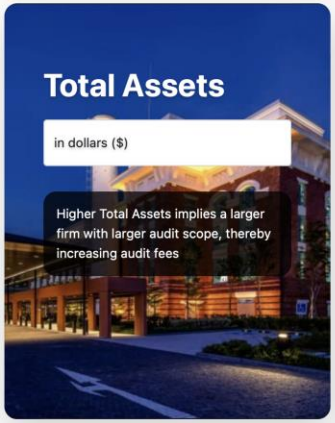
Final Regression Model Results

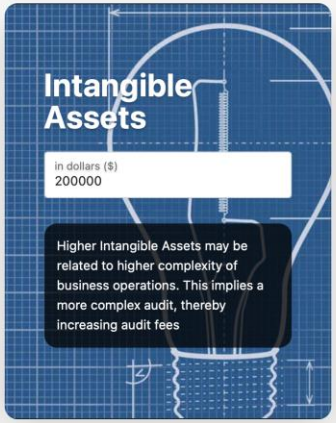
RMSE	0.5652
Adjusted R^2	0.8462
MAPE	3.2724
VIF	All <10
Significant? (p-value < 0.05)	Yes

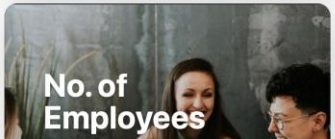
WebTool for Auditors!

https://hengweishin.wixsite.com/auditfees

Audit Fees Factors

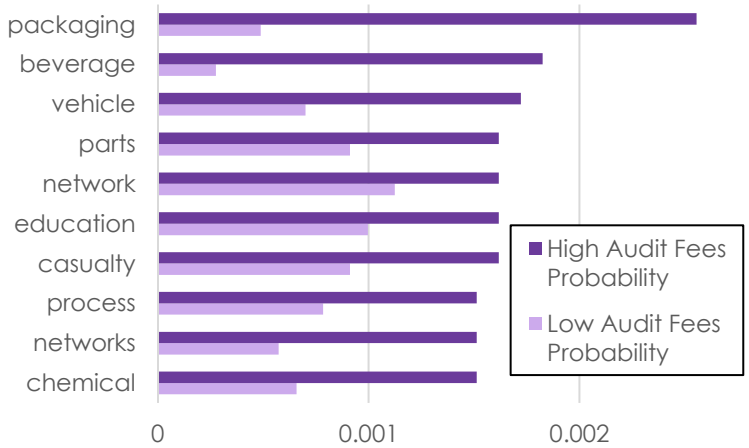






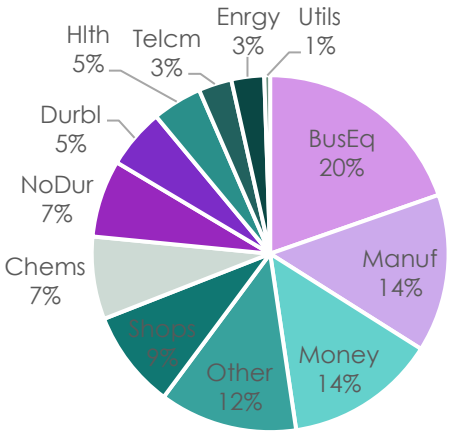
Text Analysis Results

Conditional Probability for Top 10 words of High Audit Fees



If these words are in the business description of the company, the company is likely to have higher audit fees

Breakdown of industries with busdesc in Top 10 words of High Audit Fees

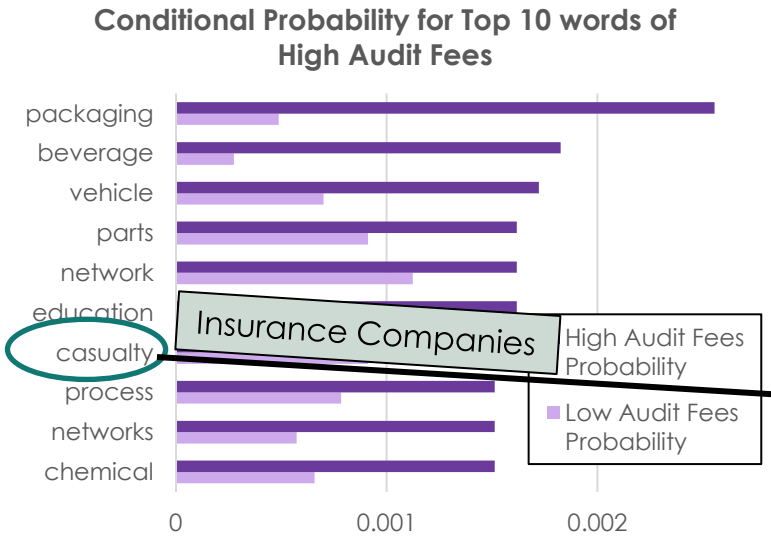


except money industry

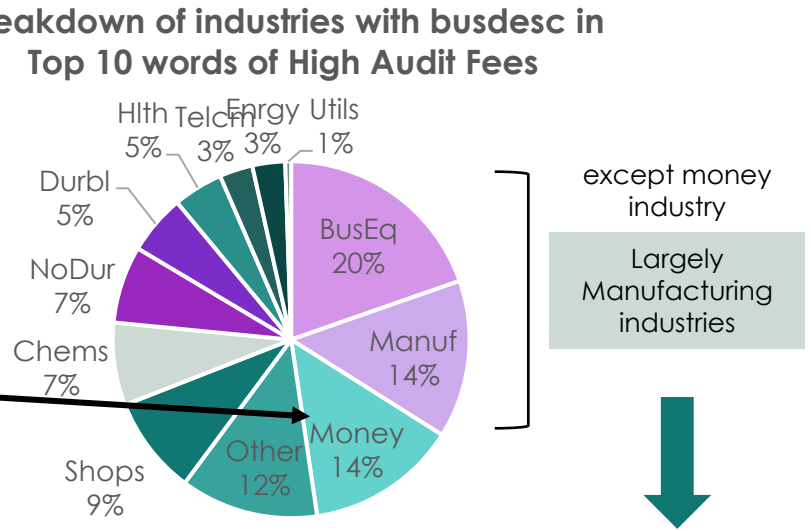
Largely Manufacturing industries

Might be due to the higher complexity of these industries. i.e. there are many stages to their production; many factors to look at

Text Analysis Results

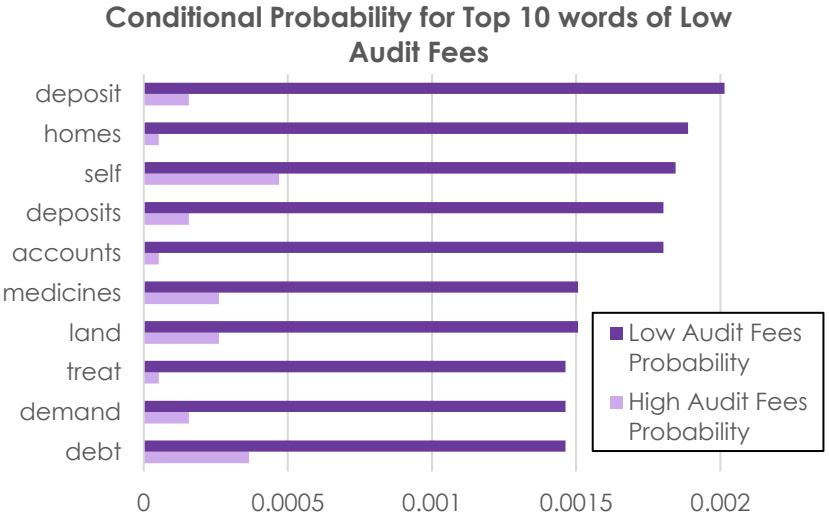


If these words are in the business description of the company, the company is likely to have higher audit fees

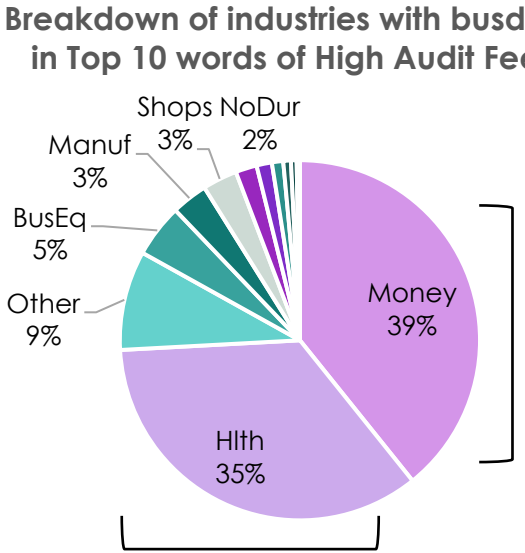


Might be due to the higher complexity of these industries. i.e. there are many stages to their production; many factors to look at

Text Analysis Results



If these words are in the business description of the company, the company is likely to have lower audit fees



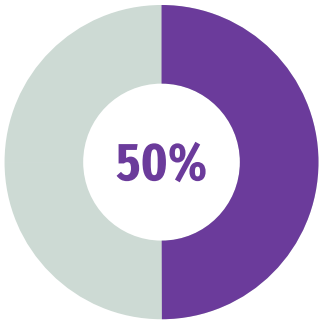
<reason why money industry lower audit fees>

<reason why money Health lower audit fees>

Possible Areas for further analysis

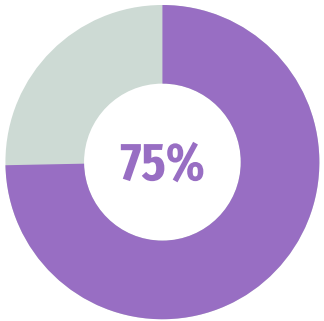
Test model on different dataset

To confirm robustness of our model



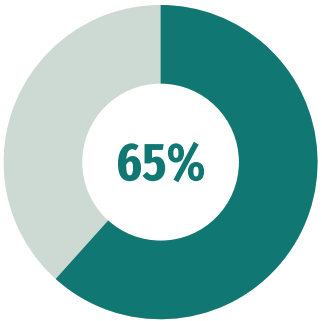
Audit fees to Total Assets Ratio

To give other variables a chance



Deep dive into other variables

Time-consuming as compared to the hypothesis approach



Time taken to perform audit

But data hard to measure and obtain

Feasibility


Thank you for listening!
Question & Ans

Appendix

Relevant Data / Further Explanations

Original Regression Model

```
Call:
lm(formula = AUDIT_FEES ~ emp + at + d_fortune + FFI12_descBusEq +
    FFI12_descChems + FFI12_descDurbl + FFI12_descEnrgy + FFI12_descHlth +
    FFI12_descManuf + FFI12_descMoney + FFI12_descNoDur + FFI12_descOther +
    FFI12_descShops + FFI12_descTelcm + FFI12_descUtils + BIG4 +
    intan, data = Audit_Reg)

Residuals:
      Min        1Q    Median        3Q       Max
-45092686  -1141402   -333491    270355  61913533

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.921e+05  1.204e+05   2.425 0.015302 *
emp              1.566e+04  4.671e+02  33.530  < 2e-16 ***
at               2.067e+01  2.131e-01  96.979  < 2e-16 ***
d_fortune        2.261e+06  7.175e+04  31.512  < 2e-16 ***
FFI12_descBusEq  4.341e+05  1.271e+05   3.415 0.000639 ***
FFI12_descChems  1.046e+06  1.824e+05   5.735 9.86e-09 ***
FFI12_descDurbl  8.708e+05  1.871e+05   4.653 3.29e-06 ***
FFI12_descEnrgy  4.927e+05  1.664e+05   2.961 0.003073 **
FFI12_descHlth  -2.923e+05  1.295e+05  -2.257 0.024018 *
FFI12_descManuf  8.314e+05  1.375e+05   6.049 1.48e-09 ***
FFI12_descMoney  1.038e+05  1.214e+05   0.855 0.392485
FFI12_descNoDur  2.061e+05  1.648e+05   1.251 0.210946
FFI12_descOther  1.084e+04  1.313e+05   0.083 0.934194
FFI12_descShops -9.435e+05  1.397e+05  -6.754 1.47e-11 ***
FFI12_descTelcm -1.422e+06  2.055e+05  -6.919 4.66e-12 ***
FFI12_descUtils         NA         NA      NA       NA
BIG4             1.748e+06  5.059e+04  34.546  < 2e-16 ***
intan            2.239e+02  3.055e+00  73.304  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3446000 on 22778 degrees of freedom
Multiple R-squared:  0.6091,    Adjusted R-squared:  0.6088
F-statistic:  2218 on 16 and 22778 DF,  p-value: < 2.2e-16
```

Log Regression Model

```
Call:
lm(formula = audit_fees_log ~ emp_log + at_log + d_fortune +
    FFI12_descBusEq + FFI12_descChems + FFI12_descDurbl + FFI12_descEnrgy +
    FFI12_descHlth + FFI12_descManuf + FFI12_descMoney + FFI12_descNoDur +
    FFI12_descOther + FFI12_descShops + FFI12_descTelcm + FFI12_descUtils +
    BIG4 + intan_log, data = Audit_train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1743 -0.3683  0.0011  0.3626  3.8337

Coefficients: (1 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.361520   0.032567 318.163  < 2e-16 ***
emp_log          0.152998   0.006194  24.701  < 2e-16 ***
at_log           0.328688   0.003687  89.144  < 2e-16 ***
d_fortune        0.064816   0.015185   4.268 1.98e-05 ***
FFI12_descBusEq  0.601533   0.027574  21.815  < 2e-16 ***
FFI12_descChems  0.592246   0.037285  15.884  < 2e-16 ***
FFI12_descDurbl  0.484467   0.038452  12.599  < 2e-16 ***
FFI12_descEnrgy  0.408597   0.032886  12.425  < 2e-16 ***
FFI12_descHlth   0.596120   0.027826  21.423  < 2e-16 ***
FFI12_descManuf  0.600727   0.029116  20.632  < 2e-16 ***
FFI12_descMoney -0.061057   0.024332  -2.509   0.0121 *
FFI12_descNoDur  0.367825   0.034323  10.716  < 2e-16 ***
FFI12_descOther  0.310686   0.027597  11.258  < 2e-16 ***
FFI12_descShops  0.152942   0.029494   5.185 2.18e-07 ***
FFI12_descTelcm  0.383178   0.040722   9.410  < 2e-16 ***
FFI12_descUtils        NA         NA      NA       NA
BIG4             0.732246   0.012125  60.392  < 2e-16 ***
intan_log        0.061357   0.002341  26.206  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16
```

Forward

```
Call:
lm(formula = audit_fees_log ~ emp_log + at_log + d_fortune + 
    FFI12_descBusEq + FFI12_descChems + FFI12_descDurbl + FFI12_descEnrgy + 
    FFI12_descHlth + FFI12_descManuf + FFI12_descMoney + FFI12_descNoDur + 
    FFI12_descOther + FFI12_descShops + FFI12_descTelcm + BIG4 + 
    intan_log, data = Audit_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1743 -0.3683  0.0011  0.3626  3.8337 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     10.361520   0.032567 318.163  < 2e-16 ***
emp_log          0.152998   0.006194  24.701  < 2e-16 ***
at_log           0.328688   0.003687  89.144  < 2e-16 ***
d_fortune        0.064816   0.015185   4.268 1.98e-05 ***
FFI12_descBusEq  0.601533   0.027574  21.815  < 2e-16 ***
FFI12_descChems  0.592246   0.037285  15.884  < 2e-16 ***
FFI12_descDurbl  0.484467   0.038452  12.599  < 2e-16 ***
FFI12_descEnrgy  0.408597   0.032886  12.425  < 2e-16 ***
FFI12_descHlth   0.596120   0.027826  21.423  < 2e-16 ***
FFI12_descManuf  0.600727   0.029116  20.632  < 2e-16 ***
FFI12_descMoney -0.061057   0.024332  -2.509   0.0121 *  
FFI12_descNoDur  0.367825   0.034323  10.716  < 2e-16 ***
FFI12_descOther  0.310686   0.027597  11.258  < 2e-16 ***
FFI12_descShops  0.152942   0.029494   5.185 2.18e-07 ***
FFI12_descTelcm  0.383178   0.040722   9.410  < 2e-16 ***
BIG4             0.732246   0.012125  60.392  < 2e-16 ***
intan_log        0.061357   0.002341  26.206  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441 
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16
```

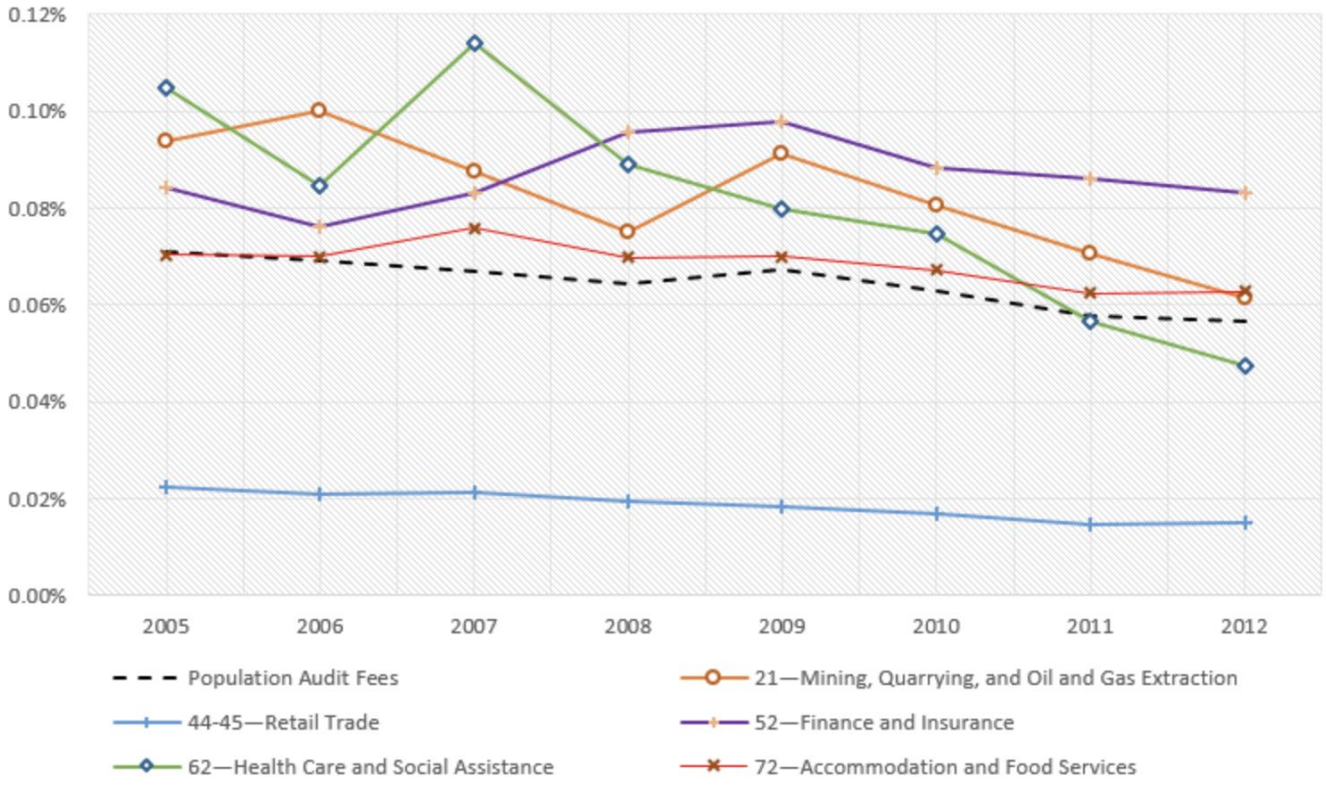
Backward

```
Call:
lm(formula = audit_fees_log ~ emp_log + at_log + d_fortune + 
    FFI12_descBusEq + FFI12_descChems + FFI12_descDurbl + FFI12_descEnrgy + 
    FFI12_descHlth + FFI12_descManuf + FFI12_descMoney + FFI12_descNoDur + 
    FFI12_descOther + FFI12_descShops + FFI12_descTelcm + BIG4 + 
    intan_log, data = Audit_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1743 -0.3683  0.0011  0.3626  3.8337 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     10.361520   0.032567 318.163  < 2e-16 ***
emp_log          0.152998   0.006194  24.701  < 2e-16 ***
at_log           0.328688   0.003687  89.144  < 2e-16 ***
d_fortune        0.064816   0.015185   4.268 1.98e-05 ***
FFI12_descBusEq  0.601533   0.027574  21.815  < 2e-16 ***
FFI12_descChems  0.592246   0.037285  15.884  < 2e-16 ***
FFI12_descDurbl  0.484467   0.038452  12.599  < 2e-16 ***
FFI12_descEnrgy  0.408597   0.032886  12.425  < 2e-16 ***
FFI12_descHlth   0.596120   0.027826  21.423  < 2e-16 ***
FFI12_descManuf  0.600727   0.029116  20.632  < 2e-16 ***
FFI12_descMoney -0.061057   0.024332  -2.509   0.0121 *  
FFI12_descNoDur  0.367825   0.034323  10.716  < 2e-16 ***
FFI12_descOther  0.310686   0.027597  11.258  < 2e-16 ***
FFI12_descShops  0.152942   0.029494   5.185 2.18e-07 ***
FFI12_descTelcm  0.383178   0.040722   9.410  < 2e-16 ***
BIG4             0.732246   0.012125  60.392  < 2e-16 ***
intan_log        0.061357   0.002341  26.206  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441 
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16

> Audit_Reg_Back_Pred <- predict(Audit_Reg_Back, Audit_test)
> accuracy(Audit_Reg_Back_Pred, Audit_test$audit_fees_log)
                 ME       RMSE        MAE         MPE      MAPE
Test set 0.003865102 0.5677058 0.4459143 -0.1492596 3.290417
```

Stepwise

```
Call:
lm(formula = audit_fees_log ~ emp_log + at_log + d_fortune + 
    FFI12_descBusEq + FFI12_descChems + FFI12_descDurbl + FFI12_descEnrgy + 
    FFI12_descHlth + FFI12_descManuf + FFI12_descMoney + FFI12_descNoDur + 
    FFI12_descOther + FFI12_descShops + FFI12_descTelcm + BIG4 + 
    intan_log, data = Audit_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1743 -0.3683  0.0011  0.3626  3.8337 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     10.361520   0.032567 318.163  < 2e-16 ***
emp_log          0.152998   0.006194  24.701  < 2e-16 ***
at_log           0.328688   0.003687  89.144  < 2e-16 ***
d_fortune        0.064816   0.015185   4.268 1.98e-05 ***
FFI12_descBusEq  0.601533   0.027574  21.815  < 2e-16 ***
FFI12_descChems  0.592246   0.037285  15.884  < 2e-16 ***
FFI12_descDurbl  0.484467   0.038452  12.599  < 2e-16 ***
FFI12_descEnrgy  0.408597   0.032886  12.425  < 2e-16 ***
FFI12_descHlth   0.596120   0.027826  21.423  < 2e-16 ***
FFI12_descManuf  0.600727   0.029116  20.632  < 2e-16 ***
FFI12_descMoney -0.061057   0.024332  -2.509   0.0121 *  
FFI12_descNoDur  0.367825   0.034323  10.716  < 2e-16 ***
FFI12_descOther  0.310686   0.027597  11.258  < 2e-16 ***
FFI12_descShops  0.152942   0.029494   5.185 2.18e-07 ***
FFI12_descTelcm  0.383178   0.040722   9.410  < 2e-16 ***
BIG4             0.732246   0.012125  60.392  < 2e-16 ***
intan_log        0.061357   0.002341  26.206  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5672 on 15939 degrees of freedom
Multiple R-squared:  0.8442,    Adjusted R-squared:  0.8441 
F-statistic:  5399 on 16 and 15939 DF,  p-value: < 2.2e-16

> Audit_Reg_Stepwise_Pred <- predict(Audit_Reg_Stepwise, Audit_test)
> accuracy(Audit_Reg_Stepwise_Pred, Audit_test$audit_fees_log)
                 ME       RMSE        MAE         MPE      MAPE
Test set 0.003865102 0.5677058 0.4459143 -0.1492596 3.290417
```

Big 4 Audit Fees as a Percentage of Total Revenue: Comparison of NAICS Sector to Russell 3000 Population



Text Analysis Methodology

Preparing the data

library(superml)

```
cv = CountVectorizer$new(remove_stopwords =
TRUE,
                    lowercase = TRUE,
                    regex = "/^[A-Za-z]+$/",
                    max_df = 0.008)
cv_matrix = read_csv("data/word counts.csv")
```

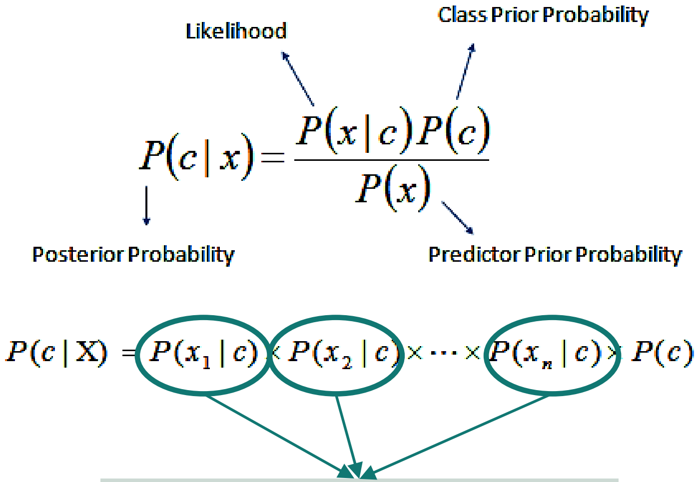
Creates a dummy variable for each unique word in the busdesc column

xylem	yacht	yale	yards	yellow	yelp	yew	yext	yield	york	yo
0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	

A snippet of the 3580 by 7830 dataframe
3580 business descriptions
7830 unique words

Utilized a technique called Naïve Bayes

library(naivebayes)



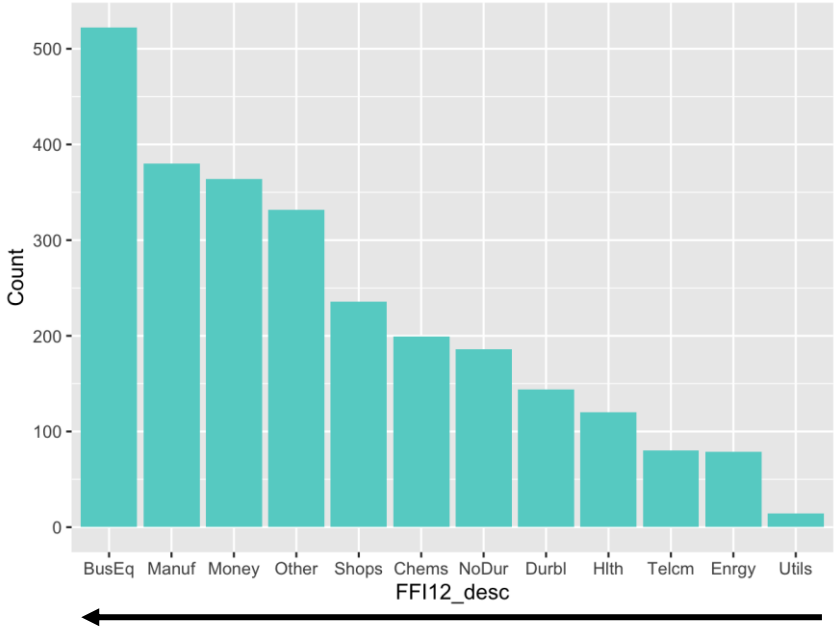
$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Calculated this probability for each word in the vocab list

Text Analysis Results – Visualisation with R

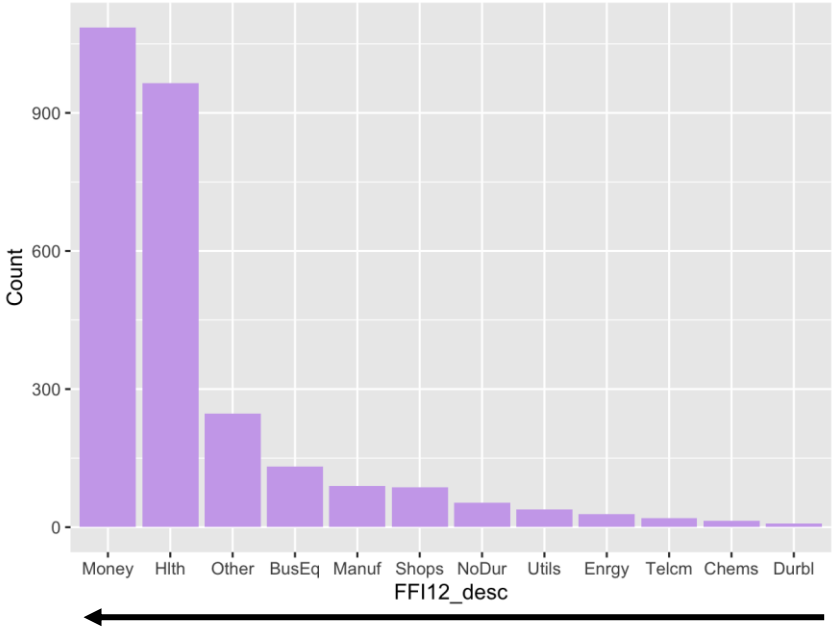
Probability of High Audit Fees by Industry



Highest Probability

Lowest Probability

Probability of Low Audit Fees by Industry



Highest Probability

Lowest Probability