DA3 – Assignment 2: Finding fast growing firms

Authors:

Saad Joiya – 2300715

Matyas Kovacs - 2405833

Executive summary

This study develops a predictive framework to identify fast-growing firms using panel data with a baseline year of 2012. To capture different growth dynamics, we define two types of fast growth: short-term (1-year) growth, which reflects explosive expansion within a year, and long-term (2-year) growth, which accounts for sustained development beyond an immediate surge (More details in Label Engineering). We employ Logit regression, LASSO, and Random Forest to classify firms as fast-growing. Our Random Forest model performs best, achieving the lowest RMSE (0.401), AUC (0.681), and expected loss (0.6118). Given the investment perspective, we prioritize recall over accuracy or precision, ensuring most fast-growing firms are correctly identified while accepting a higher false-positive rate.

To test the model's validity, we apply it separately to the Manufacturing and Services industries. Results show significant differences, at an overall level Services sector performing similarly to the full dataset (loss ≈ 0.60), overperforming on recall and underperforming on other metrics, while Manufacturing underperforms (loss ≈ 0.72) on all metrics. These findings highlight industry-specific growth patterns and suggest that tailored models may improve predictive accuracy. The study provides actionable insights for investors seeking to identify high-growth firms while balancing risk and reward.

Introduction

Fast-growing firms drive economic growth, create jobs, and foster innovation. Identifying them early can help investors, policymakers, and business leaders make smarter decisions. Amongst all these stakeholders, our analysis concentrates its focus on investors and how we can help them make wise investment decisions that reap maximum profits.

Predicting firm growth is complex and is influenced by financial performance, industry trends, and broader economic conditions. This study uses a comprehensive data-driven approach to label and predict high-growth firms and examines whether growth patterns vary across industries. By bridging machine learning techniques with firm growth analysis, this study provides actionable insights for stakeholders seeking to identify promising investment opportunities while maintaining the delicate balance between risk and reward.

Data

Data source

The dataset consists of Bisnode firm **data spanning from 2005 to 2015**, structured as panel data where each row represents a firm per given year. It includes three main types of variables: firm financials and performance, firm characteristics and lifecycle, and additional contextual variables. This comprehensive dataset captures firm financial health, growth dynamics, and leadership attributes over time, making it well-suited for predicting fast-growing firms. All features are presented in Appendix Table 7.

Label engineering

A firm is considered active if its sales exceed zero. We **include only data before 2014** to maintain relevance. **Companies** that were **active in 2012** will be analyzed. Growth is measured using 1-year (2012–2013) and 2-year (2012–2014) metrics, capturing short-term surges and long-term expansion, respectively. Sales are log-transformed and normalized by industry median for stability and comparability. Fast-growing firms are those in the **top 30% (1-year) and top 40% (2-year)** within their industry. This gives around 20% of companies as fast growing.

Feature engineering

We generate categorical variables to improve predictive accuracy. Firms younger than one year are classified as new, and firms with foreign-majority management are flagged. Negative asset values are set to zero with a corresponding flag. Normalization is applied based on relevance: sales-related variables by sales, asset-related by total assets. Non-negative accounting variables are adjusted with zero replacements and flag indicators. Outliers are winsorized for most features and categorical flags are created for CEO age. Log transformations, sales-squared term, and extreme sales flags capture nonlinear growth. Finally, unnecessary variables are dropped to streamline the dataset. (Details of raw and engineered features in appendix figure 19).

Predicting fast-growing firms

In building our predictive models, we split the dataset into a **training set (80%)** and a **holdout set (20%)** to evaluate model performance on unseen data. During model training, we employ five-fold cross-validation to ensure robustness and prevent overfitting. Model performance is initially assessed using Root Mean Squared Error (**RMSE**), and we set the scoring metric to **negative Brier score**. However, to evaluate **classification** performance, we later switch to the **ROC AUC** score to better capture the models' ability to distinguish between fast-growing and non-fast-growing firms.

Logit Models

For the Logit models, we systematically build increasingly complex models by grouping variables into different sets. We start with a basic model and progressively add more variables to capture additional firm characteristics and interactions. To enhance predictive power, we explore potential interaction effects and incorporate relevant interaction terms into our models. In total, we develop six models, each expanding on the previous one by including additional features (details in appendix 20). This step-by-step approach allows us to assess the impact of different variable groups on predicting fast-growing firms. We calculate 5-fold RMSE and AUC for all Logit models (appendix 8 & 9).

LASSO Model

For the LASSO model, we use the most complex logit model (sixth model). We begin by **standardizing all variables** so that they have a mean of 0 and a standard deviation of 1. This step is crucial because, without standardization, variables with larger scales would have larger absolute coefficients, causing LASSO to penalize them more heavily. Standardization ensures that all variables receive equal penalization, promoting stability and enhancing feature selection. We then create a range of 10 lambda and c values, perform cross-validation to **get the best lambda** and c values (lowest RMSE). The optimal values are then used to calculate 5-fold RMSE and AUC for Logit Lasso model (appendix 8 & 9).

Random Forest Model

The final model we use is the Random Forest model. In this case, we utilize the non-transformed variables (appendix 20), as Random Forest is capable of handling transformations and nonlinear relationships on its own. During the modeling process, we perform a grid search to identify the optimal hyperparameters and carry out cross-validation for each combination of parameters. To evaluate the model's performance, we first use RMSE to compare its fit with the previous models. Subsequently, we shift to using AUC to assess the model's classification ability, providing a more nuanced evaluation of its predictive performance on all 5 folds (appendix 8 & 9).

Model results

Now that we have built our models, we proceed to evaluate their average performance. The following table summarizes the results, allowing us to compare the effectiveness of Logit, LASSO, and Random Forest models in predicting fast-growing firms.

Model	Number of Coefficients	CV RMSE	CV AUC
Logit1	14	0,4093	0,6287
Logit2	35	0,4078	0,6349
Logit3	50	0,4058	0,6519
Logit4	65	0,4051	0,6572
Logit5	100	0,4052	0,6559
Logit6	187	0,4038	0,6647
LASSO	89	0,4006	0,6797
RF	n.a.	0,3986	0,6837

1. Table Model results

The results show that model performance improves with increasing complexity. As more confounders are added, Logit models see decreasing RMSE and increasing AUC, but Logits 4 and 5 perform similarly, indicating diminishing returns from quadratic and flag variables. LASSO further improves RMSE and AUC while pruning over half of the variables, enhancing interpretability. **Random Forest achieves the best RMSE and AUC, though only marginally better than LASSO**. Its ability to handle raw variables without transformations or interaction terms makes it the most efficient choice. Next, we evaluate classification performance based on threshold values.

Classification and Best Thresholds

In the context of classification and threshold selection, our primary concern is minimizing the risk of misclassifying fast-growing firms as non-fast-growing. From the perspective of an investment fund, failing to identify a truly fast-growing company is more costly than mistakenly classifying a non-fast-growing firm as fast-growing. This is because the potential profit from investing in a fast-growing firm outweighs the potential loss from investing in a misclassified firm. To reflect this imbalance, we assign a higher **cost ratio of 4:1**, meaning that misclassifying a fast-growing company is considered four times more costly than incorrectly labeling a non-fast-growing firm as fast-growing. This cost consideration will guide our selection of the optimal classification threshold.

To determine the optimal classification threshold, we analyze the expected loss for each model. By **calculating the expected loss values across different thresholds**, we create a loss plot, which helps visualize how the choice of threshold impacts misclassification costs. Additionally, we generate a ROC plot to visually examine the true positive rate and false positive rate for each model, providing insight into their overall classification performance. For reference, in the appendix we provide the distribution of predicted probabilities by Logit & Lasso models (graph 10), 5th fold Loss and ROC plot for Lasso (graph 11,12) and Random Forest (graph 13,14). At an overall level, the table below **summarizes the key results from these models**:

Model	Number of Coefficients	CV RMSE	CV AUC	CV treshold	CV expected Loss
Logit1	14	0,4093	0,6287	0,2121	0,6698
Logit2	35	0,4078	0,6349	0,2088	0,6701
Logit3	50	0,4058	0,6519	0,2063	0,6499
Logit4	65	0,4051	0,6572	0,1956	0,6399
Logit5	100	0,4052	0,6559	0,1968	0,6410
Logit6	187	0,4038	0,6647	0,1962	0,6291
LASSO	89	0,4006	0,6797	0,1993	0,6166
RF	n.a.	0,3986	0,6837	0,2077	0,6118

2. Table Model results with thresholds and exp. loss

Discussion of results

Our results show that the Random Forest model performs best across loss, RMSE, and AUC. Using it on the holdout set, it achieves **RMSE = 0.401, Loss = 0.628, and AUC = 0.681**. Its calibration curve suggests good alignment at lower probabilities but slight deviation at higher ones. Using an **optimal threshold of 0.20** we get the confusion matrix below:

	Predicted No Fast Growth	Predicted Fast Growth
Actual No Fast Growth	1489	1444
Actual Fast Growth	238	637

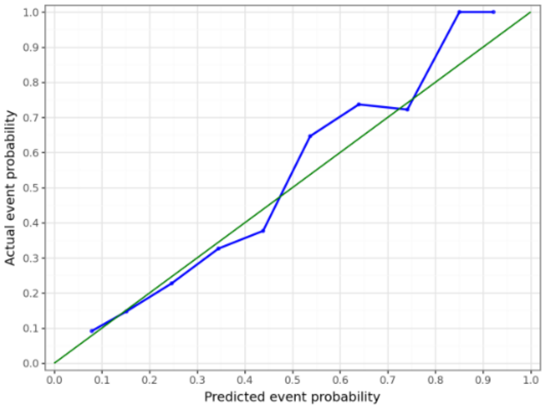
3. Table Confusion matrix Random Forest

Using an optimal threshold of 0.20, the model achieves:

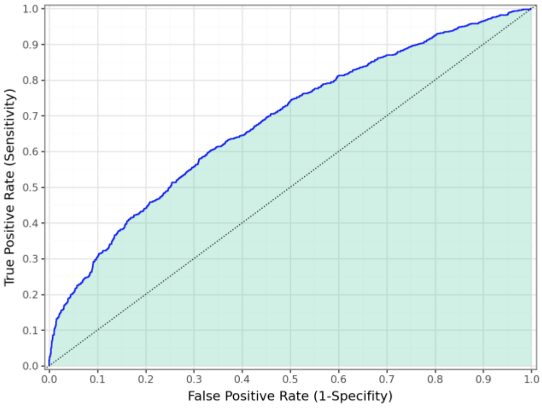
Accuracy	Recall	Precision	Specificity
55.8%	72.8%	30.6%	50.8%

4.1 Table for Qualitative and Quantitative Measure Random Forest

Prioritizing recall over precision, the model correctly identifies three out of four fast-growing firms but misclassifies two non-fast-growing firms per correct prediction. Given the investment fund's focus, missing a high-growth firm is costlier than a false positive, making this tradeoff acceptable and the model a suitable choice. You can also see the results visually in the ROC plot and Calibration Curve on the holdout-set below:



3.3 Figure – RF Calibration Curve on holdout-set



3. 4. Figure - RF ROC plot on holdout-set

Predicting fast-growing firms for separate industries

To assess the validity of our models across industries, we create two sub-datasets: Manufacturing and Services. After separating the datasets, we use only our best-performing model, Random Forest. For both industries, we first **cross-validate on work set using the same cost function** as complete data (4:1) and then **evaluate the results on the holdout subset**. The findings reveal performance differences between Manufacturing and Services subsets.

	RMSE	ROC AUC	Loss	Optimal Threshold
Manufacturing	0.419	0.612	0.721	0.198
Services	0.396	0.698	0.606	0.155

4. Figure – Performance Manufacturing vs Services on holdout

Confusion matrices and Measurements on their optimal threshold:

Manufacturing	Predicted No Fast Growth	Predicted Fast Growth
Actual No Fast Growth	275	289
Actual Fast Growth	60	117

5.1 Table Confusion matrix - Manufacturing

Services	Predicted No Fast Growth	Predicted Fast Growth
Actual No Fast Growth	723	1646
Actual Fast Growth	87	611

5.2 Table Confusion matrix - Services

	Accuracy	Recall	Precision	Specificity
Manufacturing	52.6%	62.7%	28.8%	48.0%
Services	45.1%	86.3%	27.5%	32.9%

6. Table Comparative Model Performance

Observing RMSE, AUC and Loss Services industry model outperforms the Manufacturing industry model. The Accuracy of the Services models is slightly lower than Manufacturing. However, our aim is capture as many fast-growing firms as possible (high recall) even at the expense of misclassifying some non-fast-growing firm. The recall for Services industries is much higher than Manufacturing making the trade-off seem beneficial for our intended aim. The calibration curves (appendix 15 & 17) and ROC plots (Appendix 16 & 18) further highlight these differences, with Manufacturing consistently underperforming compared to Services.

Summary

Our study predicts fast-growing firms using panel data (2005–2015) with a 2012 baseline year, defining fast growing firms using short-term (1-year) and long-term (2-year) growth. Leveraging different Logit, Lasso and Random Forest models, we observed that **Random Forest delivers the best predictive power given our customized loss function**. Our loss function emphasizes on maximizing recall at the expense of other metrics, given an investor's perspective —**ensuring high-growth firms are identified**, even if it means accepting more false positives.

At a granular level, Industry-specific patterns show stark contrasts: while Services sector outperforms the overall dataset's performance in terms of recall it underperforms in terms of accuracy, specificity and precision (aligning with aim/cost function). Manufacturing sector underperforms in all metrics compared to overall dataset. This highlights the **importance of industry specific tailored predictive models** to provide actionable insights for investors aiming to navigate the delicate balance between risk and opportunity in high-growth firm selection.

Appendix

Complete code for this analysis can be found at: ***https://github.com/joiya-saad/Data-Analysis-3/blob/9c6fa9270f002005d560777aaf038523639ab5c6/Assignment%202/Assignment%202.ipynb***

Variable	Type	Description
year	Integer	Year of observation
comp_id	Integer	Unique firm identifier
curr_assets	Float	Current assets
curr_liab	Float	Current liabilities
fixed_assets	Float	Fixed assets
intang_assets	Float	Intangible assets
inventories	Float	Inventories
liq_assets	Float	Liquid assets
tang_assets	Float	Tangible assets
share_eq	Float	Shareholder equity
subscribed_cap	Float	Subscribed capital
sales	Float	Revenue (sales)
material_exp	Float	Material expenses
personnel_exp	Float	Employee costs
extra_exp	Float	Extraordinary expenses
extra_inc	Float	Extraordinary income
profit_loss_year	Float	Net profit/loss
inc_bef_tax	Float	Income before tax
extra_profit_loss	Float	Extraordinary profits or losses
ln_sales	Float	Log-transformed sales
sales_mil	Float	Revenue in millions
sales_mil_log	Float	Log of revenue in millions
past_d1_sales_mil_log	Float	Past revenue log transformation
founded_year	Integer	Year of founding
age	Integer	Firm age
exit_year	Integer	Year of exit (if applicable)
fast_growth	Boolean	High-growth firm indicator
balsheet_flag	Boolean	Balance sheet available flag
balsheet_length	Integer	Balance sheet length
balsheet_notfullyear	Boolean	Incomplete balance sheet indicator
ceo_count	Integer	Number of CEOs over time
foreign_management	Boolean	Foreign executive indicator
gender_m	Boolean	CEO gender (Male)
female	Boolean	CEO is female indicator
urban_m	Boolean	Urban classification
m_region_loc	Integer	Regional location code
foreign	Boolean	Foreign-owned firm indicator

origin	String	Ownership origin
ind2_cat	Integer	Industry category
status_alive	Boolean	Firm is still active indicator
new	Boolean	New firm indicator
flag_asset_problem	Boolean	Financial distress flag
birth_year	Integer	CEO birth year
inoffice_days	Integer	CEO tenure in days

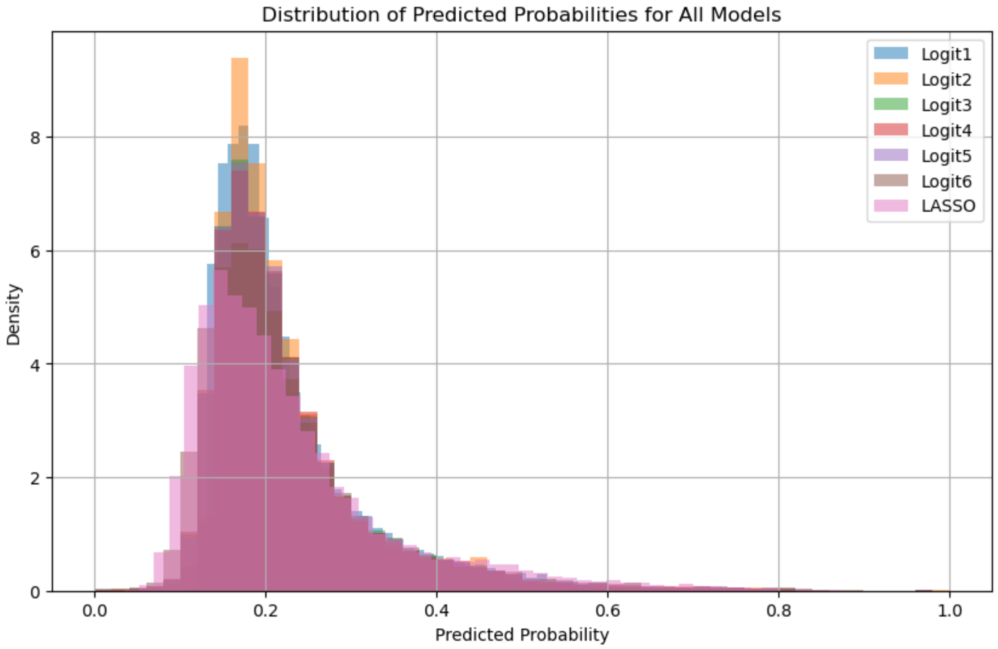
7. Table Variables of the dataset

	Logit1	Logit2	Logit3	Logit4	Logit5	Logit6	LASSO	RF
Fold 1	0.400849	0.400212	0.396436	0.396842	0.396841	0.394435	0.392934	0.390106
Fold 2	0.408557	0.407407	0.405084	0.405336	0.405456	0.404393	0.400551	0.397370
Fold 3	0.408647	0.408117	0.406804	0.405608	0.405480	0.403804	0.398684	0.400340
Fold 4	0.415422	0.413475	0.412037	0.412678	0.411971	0.410115	0.407776	0.406038
Fold 5	0.412782	0.409769	0.408534	0.405132	0.406411	0.406320	0.403112	0.399188

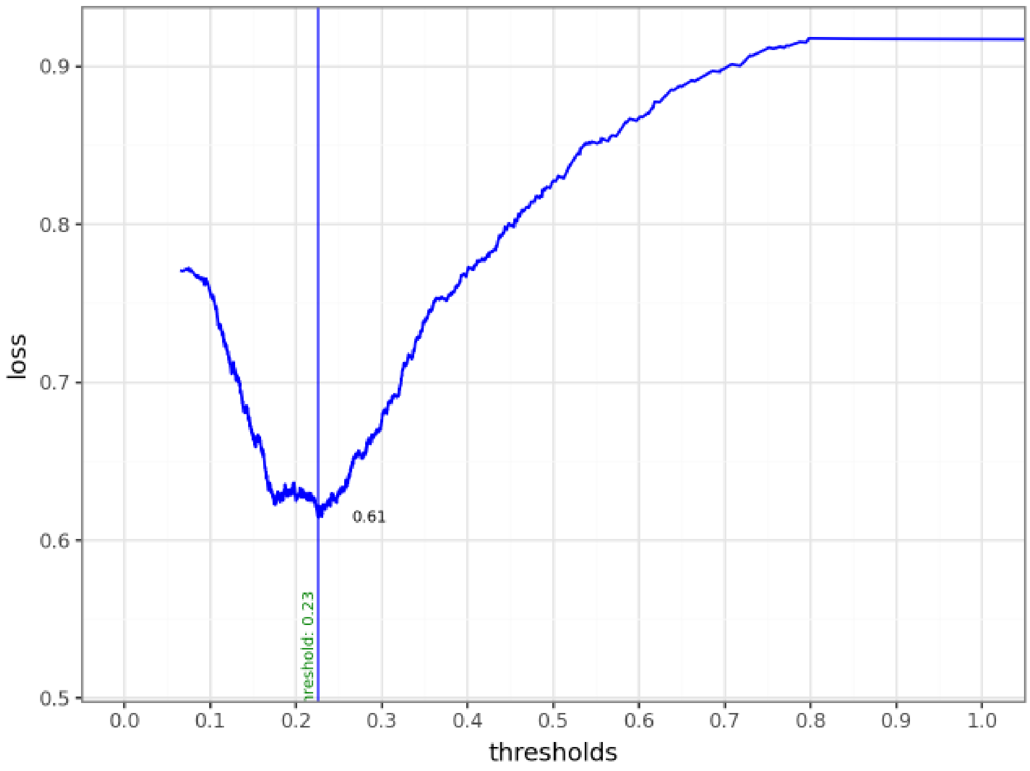
8. Table Cross validation results RMSE

	Logit1	Logit2	Logit3	Logit4	Logit5	Logit6	LASSO	RF
Fold 1	0.643045	0.648923	0.682974	0.681239	0.679619	0.692530	0.699133	0.703585
Fold 2	0.627084	0.622768	0.644354	0.643099	0.641603	0.655008	0.680064	0.684744
Fold 3	0.637804	0.643534	0.655161	0.664727	0.665623	0.672247	0.686674	0.680719
Fold 4	0.612857	0.626029	0.637533	0.632360	0.637960	0.645448	0.654343	0.662753
Fold 5	0.622955	0.633044	0.639542	0.664613	0.654565	0.658112	0.678435	0.686749

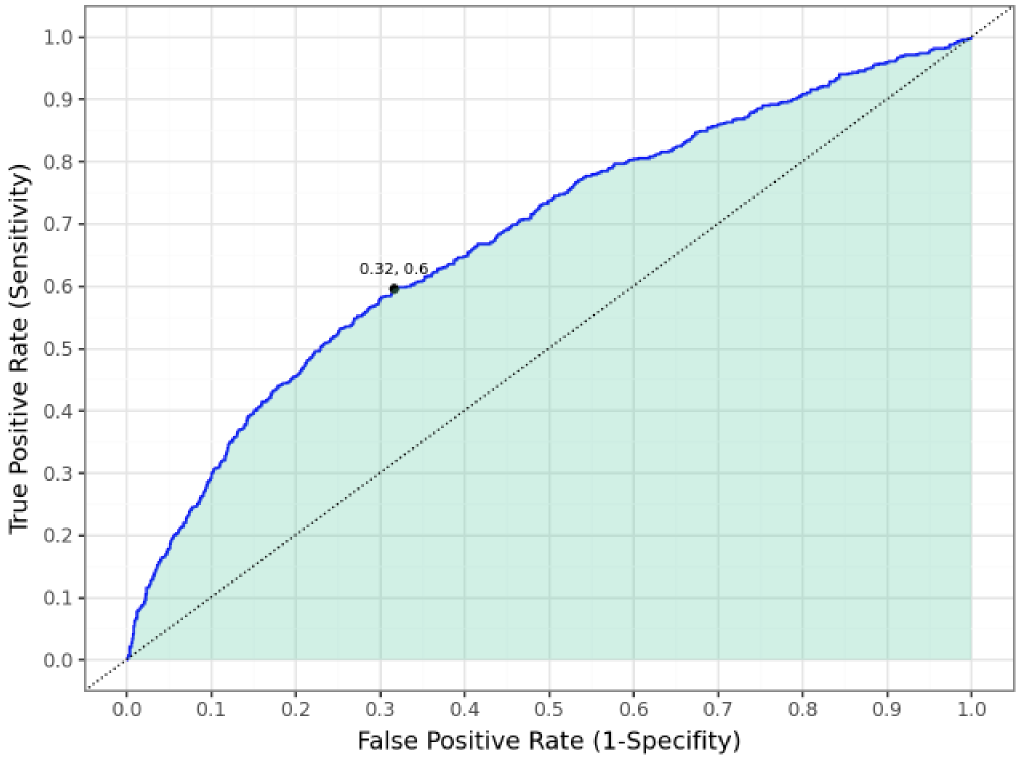
9. Table Cross validation results AUC



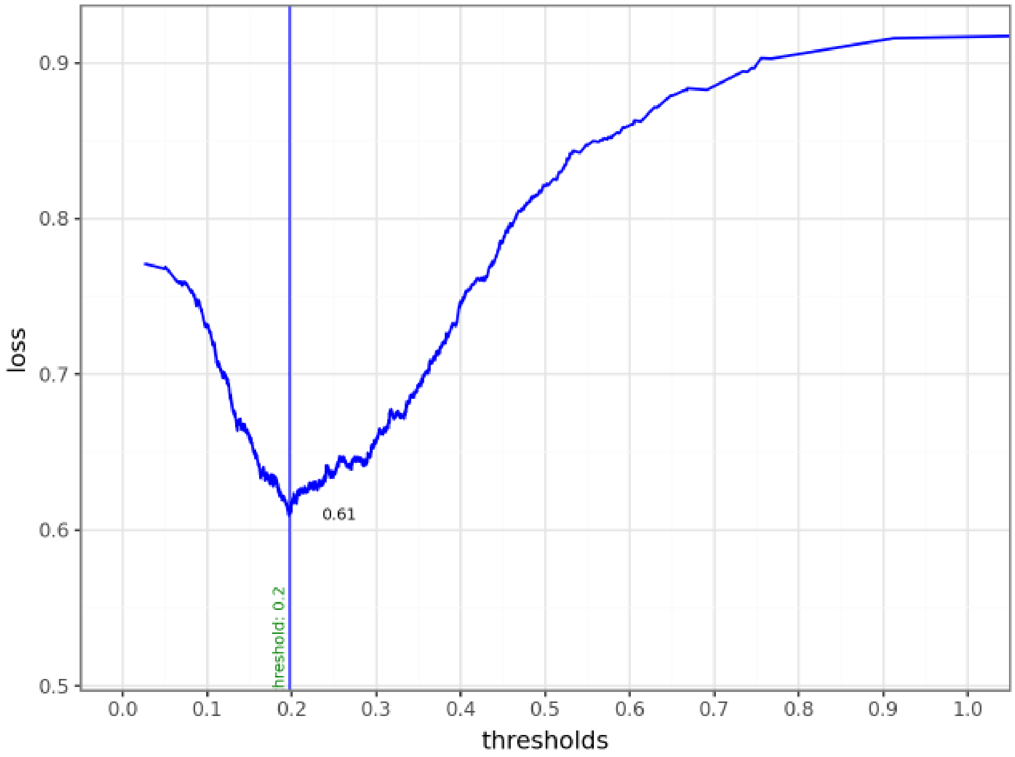
10. Figure Distribution of predicted probabilities for Logits and Lasso models



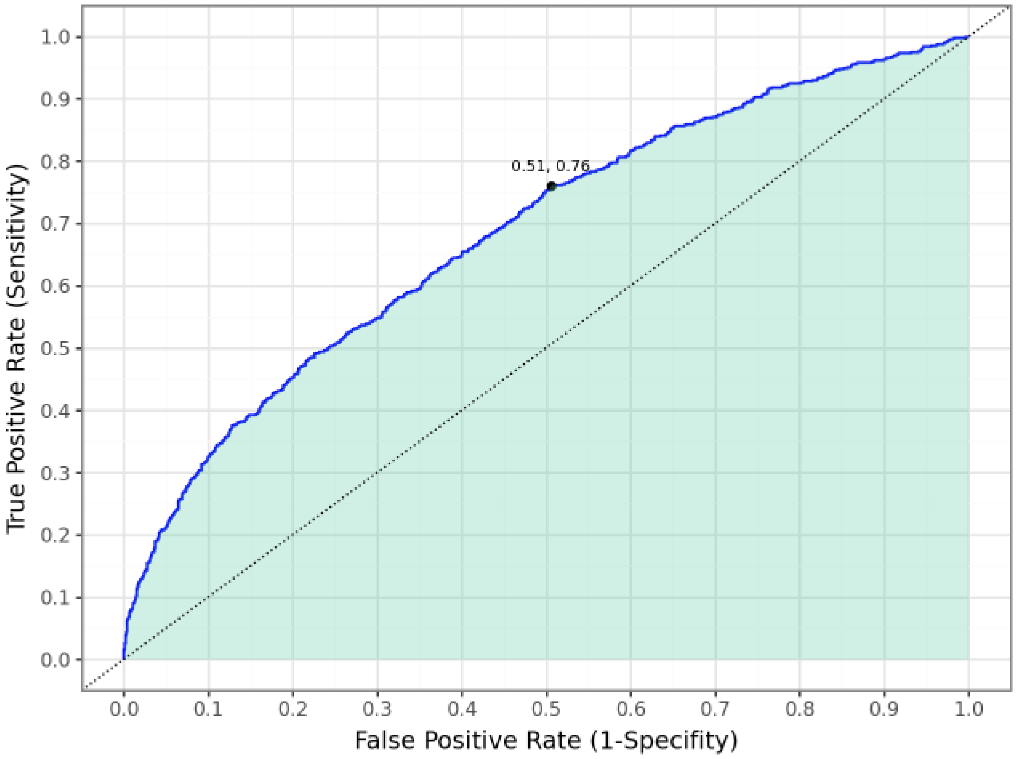
11. Figure LASSO Loss plot based on the threshold for 5th fold



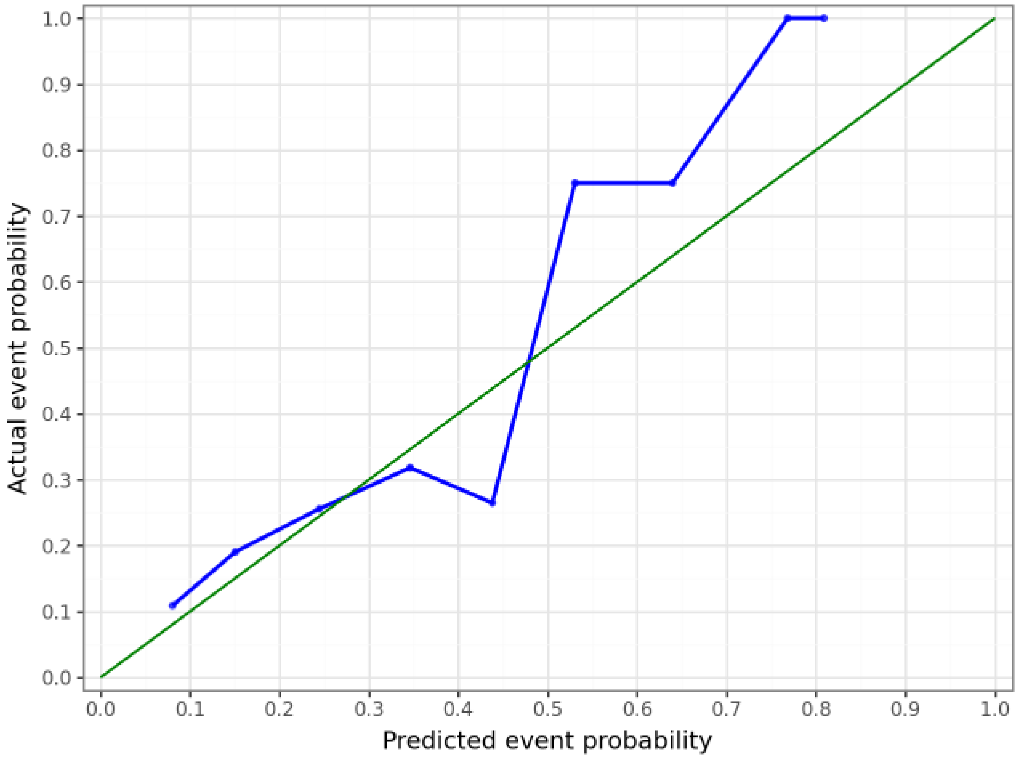
12. Figure LASSO ROC plot on 5th fold



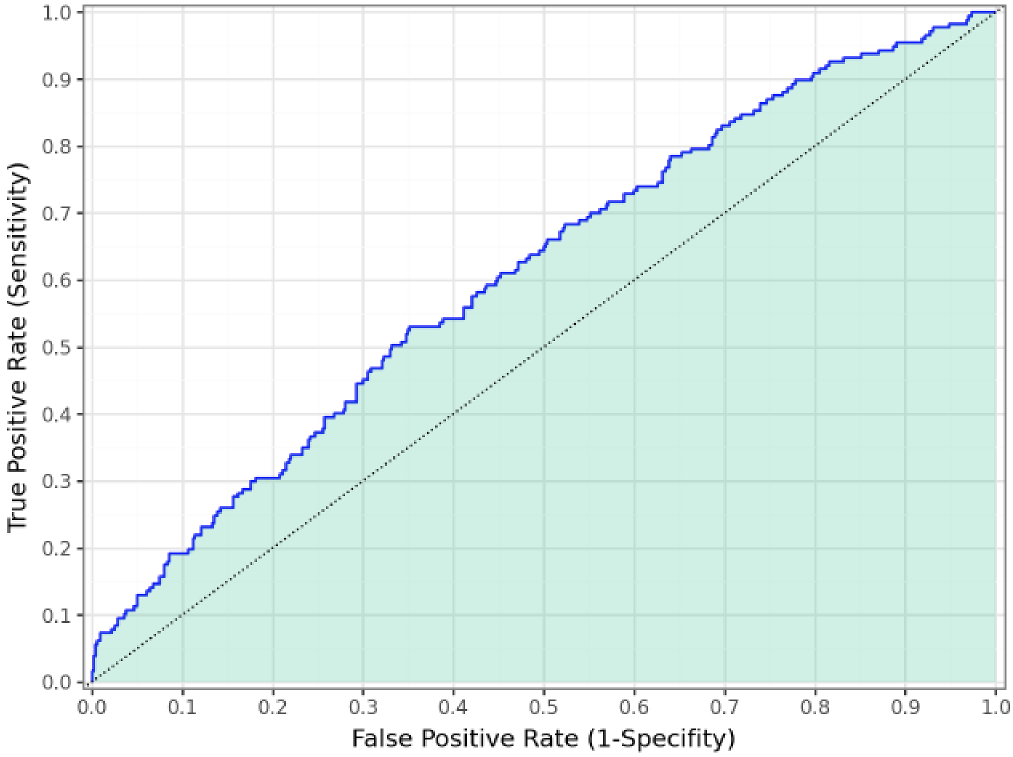
13. Figure Random Forest Loss function in 5th fold



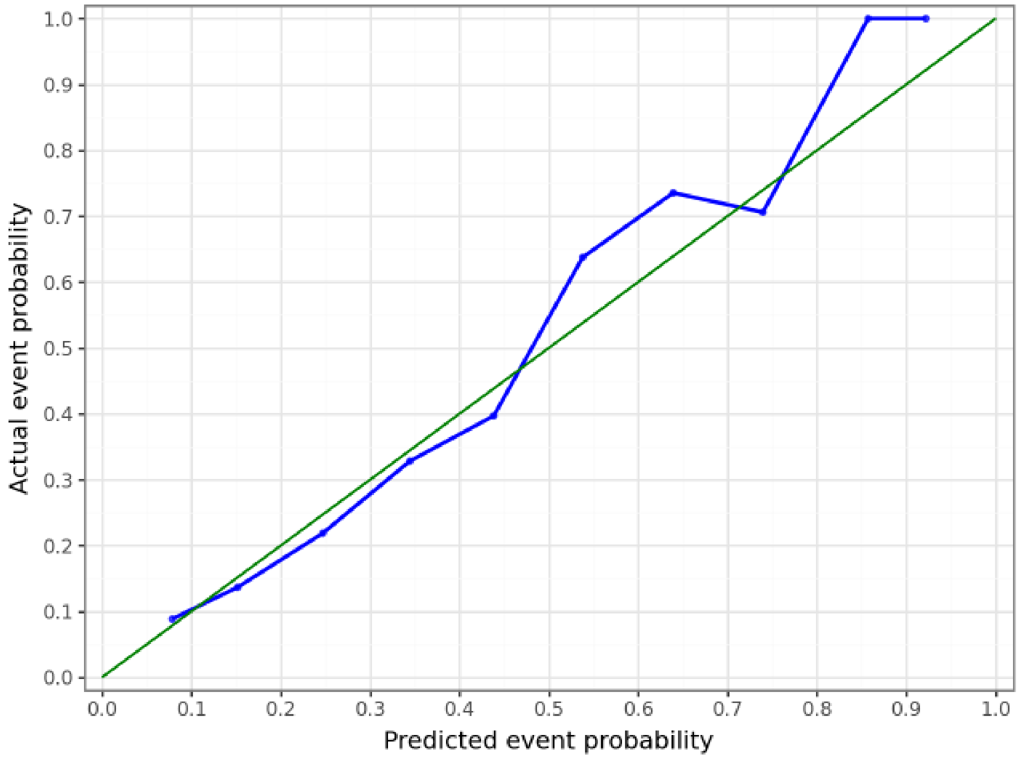
14. Figure Random Forest ROC plot on 5th fold



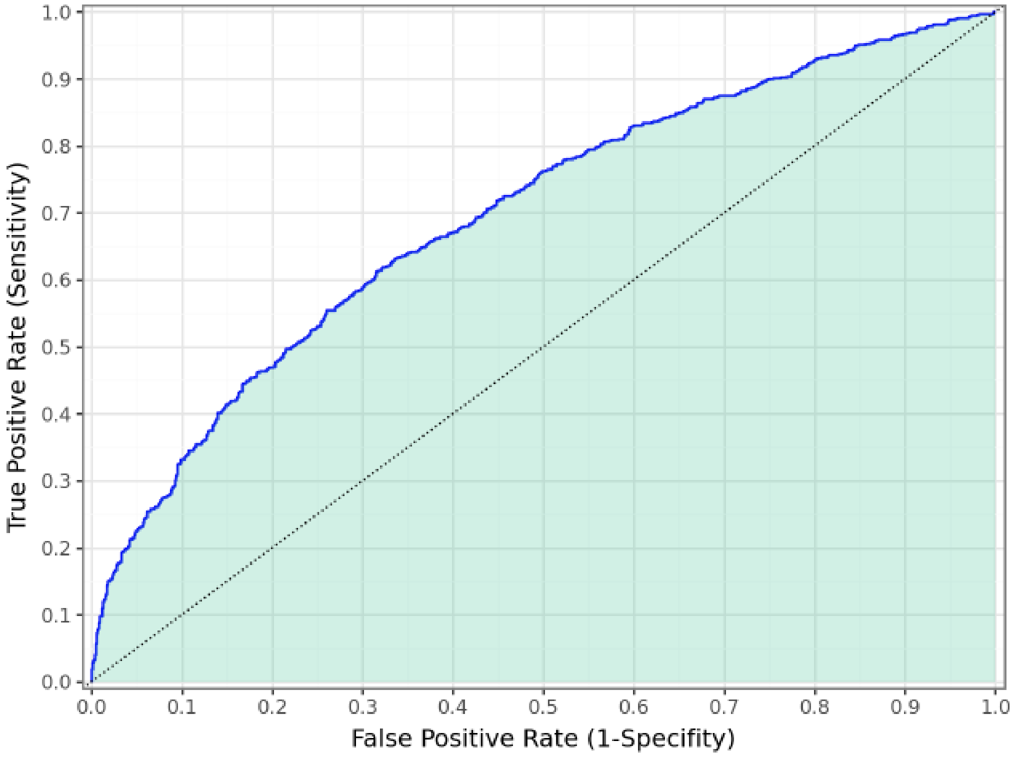
15. Figure Calibration Curve of Random Forest model on Manufacturing



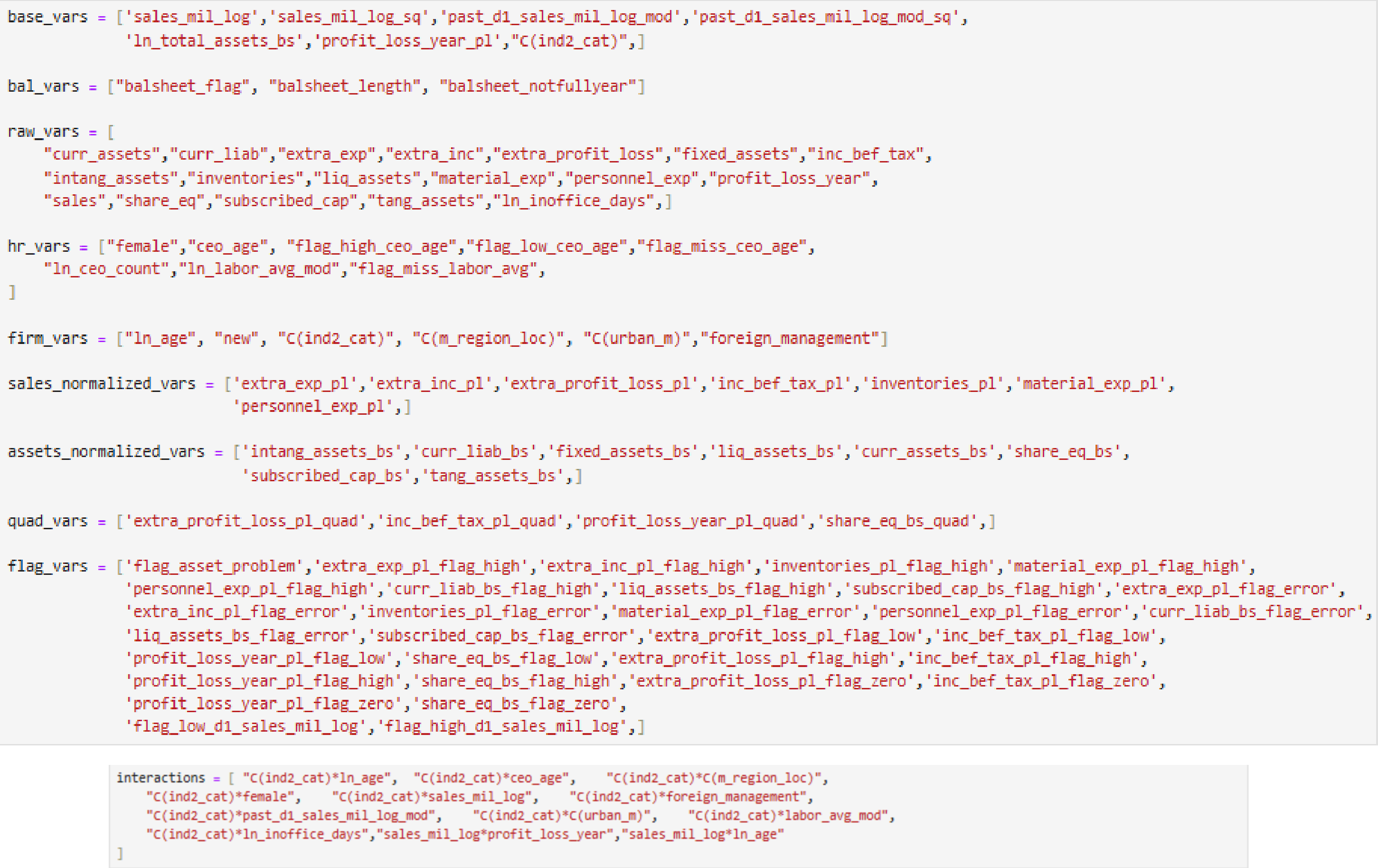
16. Figure ROC plot of Random Forest model on Manufacturing



17. Figure Calibration Curve of Random Forest model on Services



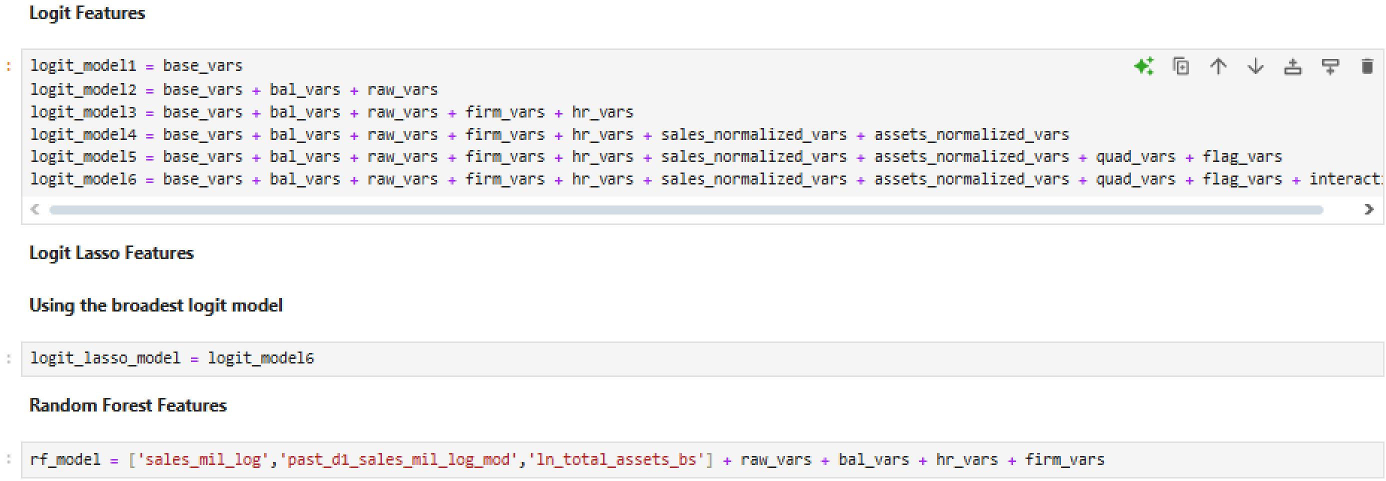
18. Figure ROC plot of Random Forest model on Services



```
base_vars = ['sales_mil_log','sales_mil_log_sq','past_d1_sales_mil_log_mod','past_d1_sales_mil_log_mod_sq',
             'ln_total_assets_bs','profit_loss_year_pl',"C(ind2_cat)",]

bal_vars = ["balsheet_flag", "balsheet_length", "balsheet_notfullyear"]

raw_vars = [
    "curr_assets","curr_liab","extra_exp","extra_inc","extra_profit_loss","fixed_assets","inc_bef_tax",
    "intang_assets","inventories","liq_assets","material_exp","personnel_exp","profit_loss_year",
    "sales","share_eq","subscribed_cap","tang_assets","ln_inoffice_days",]

hr_vars = ["female","ceo_age", "flag_high_ceo_age","flag_low_ceo_age","flag_miss_ceo_age",
    "ln_ceo_count","ln_labor_avg_mod","flag_miss_labor_avg",
]

firm_vars = ["ln_age", "new", "C(ind2_cat)", "C(m_region_loc)", "C(urban_m)","foreign_management"]

sales_normalized_vars = ['extra_exp_pl','extra_inc_pl','extra_profit_loss_pl','inc_bef_tax_pl','inventories_pl','material_exp_pl',
                         'personnel_exp_pl',]

assets_normalized_vars = ['intang_assets_bs','curr_liab_bs','fixed_assets_bs','liq_assets_bs','curr_assets_bs','share_eq_bs',
                          'subscribed_cap_bs','tang_assets_bs',]

quad_vars = ['extra_profit_loss_pl_quad','inc_bef_tax_pl_quad','profit_loss_year_pl_quad','share_eq_bs_quad',]

flag_vars = ['flag_asset_problem','extra_exp_pl_flag_high','extra_inc_pl_flag_high','inventories_pl_flag_high','material_exp_pl_flag_high',
             'personnel_exp_pl_flag_high','curr_liab_bs_flag_high','liq_assets_bs_flag_high','subscribed_cap_bs_flag_high','extra_exp_pl_flag_error',
             'extra_inc_pl_flag_error','inventories_pl_flag_error','material_exp_pl_flag_error','personnel_exp_pl_flag_error','curr_liab_bs_flag_error',
             'liq_assets_bs_flag_error','subscribed_cap_bs_flag_error','extra_profit_loss_pl_flag_low','inc_bef_tax_pl_flag_low',
             'profit_loss_year_pl_flag_low','share_eq_bs_flag_low','extra_profit_loss_pl_flag_high','inc_bef_tax_pl_flag_high',
             'profit_loss_year_pl_flag_high','share_eq_bs_flag_high','extra_profit_loss_pl_flag_zero','inc_bef_tax_pl_flag_zero',
             'profit_loss_year_pl_flag_zero','share_eq_bs_flag_zero',
             'flag_low_d1_sales_mil_log','flag_high_d1_sales_mil_log',]
```

```
interactions = [ "C(ind2_cat)*ln_age",  "C(ind2_cat)*ceo_age",    "C(ind2_cat)*C(m_region_loc)",
    "C(ind2_cat)*female",    "C(ind2_cat)*sales_mil_log",    "C(ind2_cat)*foreign_management",
    "C(ind2_cat)*past_d1_sales_mil_log_mod",    "C(ind2_cat)*C(urban_m)",    "C(ind2_cat)*labor_avg_mod",
    "C(ind2_cat)*ln_inoffice_days","sales_mil_log*profit_loss_year","sales_mil_log*ln_age"
]
```

19. Complete set of Features to be used in Models



Logit Features

```
logit_model1 = base_vars
logit_model2 = base_vars + bal_vars + raw_vars
logit_model3 = base_vars + bal_vars + raw_vars + firm_vars + hr_vars
logit_model4 = base_vars + bal_vars + raw_vars + firm_vars + hr_vars + sales_normalized_vars + assets_normalized_vars
logit_model5 = base_vars + bal_vars + raw_vars + firm_vars + hr_vars + sales_normalized_vars + assets_normalized_vars + quad_vars + flag_vars
logit_model6 = base_vars + bal_vars + raw_vars + firm_vars + hr_vars + sales_normalized_vars + assets_normalized_vars + quad_vars + flag_vars + interact
```

Logit Lasso Features

Using the broadest logit model

```
logit_lasso_model = logit_model6
```

Random Forest Features

```
rf_model = ['sales_mil_log','past_d1_sales_mil_log_mod','ln_total_assets_bs'] + raw_vars + bal_vars + hr_vars + firm_vars
```

20. Features to be used in Different Models