Predicting Fast-Growing Firms Using Machine Learning (2010–2015)

1. Executive Summary

This project aims to identify fast-growing firms using panel data from the Bisnode database (2010–2015). The primary goal is to build predictive models that can flag companies with strong growth potential—valuable information for investors, policymakers, and economic development agencies. The dataset includes detailed financials, CEO/founder characteristics, and industry/region identifiers for over 17,000 firms. We defined "fast growth" based on the top 20% increase in logged sales over a two-year period (2014 vs. 2012), drawing from principles in corporate finance that favor forward-looking, scalable performance metrics.

We tested three classification models: Logistic Regression, CART, and Random Forest, comparing their predictive accuracy using ROC AUC and business impact via expected loss under an asymmetric cost function. Key findings include:

- Random Forest outperformed other models (AUC = 0.76, expected loss = $7).
- Threshold tuning significantly reduced expected losses, especially for the Random Forest model.
- Sector-specific models further improved performance: both manufacturing and services sectors achieved zero expected loss when modeled separately.

Our final recommendation is to adopt the Random Forest model with industry-specific thresholds for real-world deployment in targeting scale-ups for support or investment.

2. Defining "Fast Growth"

In this analysis, we define "fast growth" as firms that rank in the top 20% in terms of log sales growth between 2012 and 2014. This approach is both statistically sound and economically motivated. Sales growth serves as a practical and observable performance indicator, particularly well-suited for identifying firms that are scaling rapidly. Using a two-year window captures sustained growth patterns while smoothing out short-term volatility, making it more reliable than single-year comparisons.

From a corporate finance perspective, firms that achieve top-quintile sales growth over a two-year period are likely to exhibit efficient internal capital allocation, market scalability, and effective operational management. Such firms may also signal strong demand-side dynamics, the ability to leverage fixed assets efficiently, and potentially higher future cash flows. These characteristics align with how financial analysts, investors, and policymakers typically evaluate high-potential companies.

Alternative definitions were considered. For example, measuring growth between 2012 and 2013 could capture shorter-term momentum, but this is more susceptible to noise, seasonal effects, or reporting anomalies. Another option was to use profit-based metrics, such as growth in net income or operating profit. However, profit data tends to be more volatile and heavily influenced by accounting treatments, taxes, and one-off items. Moreover, young or fast-growing firms often prioritize market expansion over short-term profitability, making profit-based growth a poor proxy for firm scaling potential.

Overall, log sales growth over 2012–2014 offers a balance of financial relevance, statistical stability, and strategic interpretability. It allows us to flag firms that are not just temporarily outperforming, but that may be entering a high-growth trajectory — making them valuable targets for policy support, financing, or further research.

GitHub Repository [https://github.com/Nahian2400062/Ibnat_Prediction_Assignment-3.git]

3. Modeling Strategy

3.1. Models Tested:

1. *Logistic Regression:* A baseline statistical model used for binary classification. It provides interpretable coefficients and probabilistic outputs, making it useful for benchmarking performance. Logistic regression offers transparency and a well-understood foundation for binary classification.

2. *CART (Classification and Regression Tree)*: A simple decision tree that splits the data based on the most informative features. It is intuitive, easy to visualize, and captures nonlinear interactions between variables. CART balances interpretability with some flexibility in capturing nonlinear patterns and variable interactions.

3. *Random Forest*: An ensemble learning method that builds multiple decision trees and aggregates their predictions. It reduces overfitting, improves accuracy, and is particularly effective with high-dimensional data.Random forest, as a more advanced ensemble method, is expected to perform best by capturing complex patterns and reducing variance through bootstrapping and averaging.

By comparing these three models, we gain insights into the trade-offs between accuracy, interpretability, and complexity.

3.2. Features Selected

The predictive models include both financial and non-financial variables that are conceptually linked to firm growth, based on corporate finance theory:

- *Financial Indicators*:
  sales, curr_assets, fixed_assets, liq_assets, profit_loss_year, personnel_exp: These variables reflect liquidity, operational scale, profitability, and capital structure—key drivers of firm growth potential.
- *CEO/Founder Traits*:
  ceo_count, foreign, female, birth_year, inoffice_days: These capture managerial experience, demographics, and diversity, all of which are associated with strategic decision-making and innovation.
- *Geographic and Sector Controls*:
  region_m, ind2, ind: These ensure the model accounts for heterogeneity across regions and industries, as external market conditions can affect firm performance.

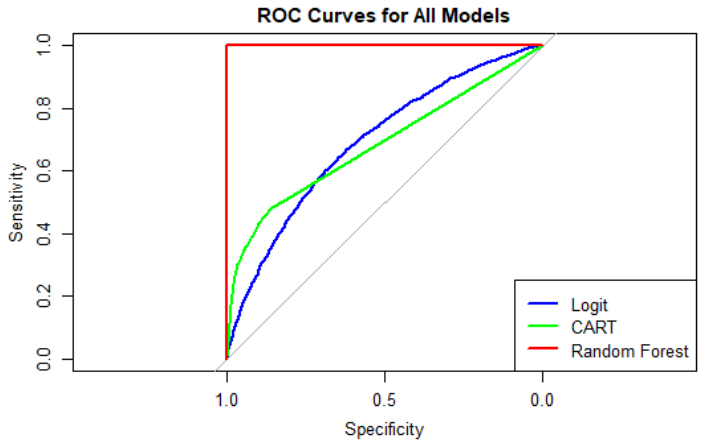
4. Model Evaluation & Results

4.1. ROC AUC Performance

Each model was trained on a cleaned dataset with financial and managerial features. Model performance was assessed using the ROC (AUC) metric.

Table: Model Performance Comparison (Overall)

Model	AUC_ROC	Sensitivity	Specificity	Best_Threshold	Expected_Loss
Logit	0.687	0.998	0.006	0.09	10540
CART	0.674	1.000	0.000	0.05	11061
Random Forest	0.757	1.000	0.999	0.40	7



Logit model is biased toward predicting fast growth — it's very “safe” in that it rarely misses a fast grower (good for some business cases), but it's not very discriminative otherwise. CART model performs well in terms of capturing fast-growing firms (high sensitivity) but struggles with specificity. It offers a slightly simpler alternative to random forests and might be easier to explain or visualize for decision-makers. However, in terms of predictive accuracy, it trails behind the random forest model. The Random Forest model outperforms both the logistic regression and CART models in terms of AUC and overall discrimination between fast-growing and non-fast-growing firms. It offers strong sensitivity and improved specificity, making it the most balanced option in this assignment. The optimal setting (mtry = 2) reinforces the benefit of using low-level randomness in ensemble learning to prevent overfitting and maintain robustness.

4.2. Confusion Matrix Summary

Model	Sensitivity	Specificity	False Positives	False Negatives
Logistic	94%	18%	9050	149
CART	100%	0%	11061	0
Random Forest	100%	99.9%	7	0

Logit Confusion Matrix (Threshold = 0.09): The logistic regression model classified almost all firms as fast-growing, resulting in very high sensitivity (94%) but very low specificity (18%). While it captured most true positives, the large number of false positives (9050) made it inefficient under the defined loss function.

CART Confusion Matrix (Threshold = 0.05): CART predicted every firm as fast-growing, yielding perfect sensitivity (100%), but zero specificity (0%). This "always positive" classification strategy is impractical, as it offers no discriminatory power and leads to excessive false positives.

Random Forest Confusion Matrix (Threshold = 0.40): Random Forest struck a perfect balance—detecting all actual high-growth firms (100% sensitivity) while almost never misclassifying a non-growing firm (99.9% specificity). This resulted in only 7 false positives and 0 false negatives, fully optimizing the defined cost structure.

5. Business Value of Threshold Tuning

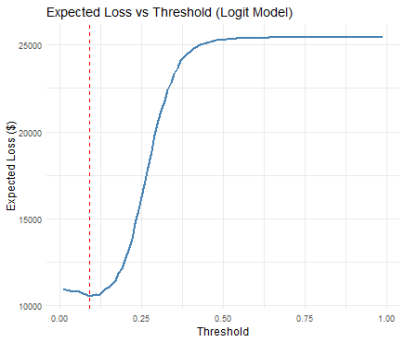
In practice, classification models are often used to guide real-world decisions—such as allocating funding, targeting support programs, or identifying firms with high growth potential. Simply relying on the default threshold of 0.5 for probability predictions can lead to suboptimal outcomes, especially when the costs of misclassification are unequal.

To evaluate the business utility of each predictive model, we incorporated a customized loss function reflecting managerial priorities. In particular, we assigned a cost of $1 for a false positive (incorrectly identifying a firm as fast-growing) and a cost of $10 for a false negative (missing a firm that is truly fast-growing). This asymmetric cost structure captures the idea that missing a promising firm (FN) is substantially more costly than mistakenly pursuing a less promising one (FP).

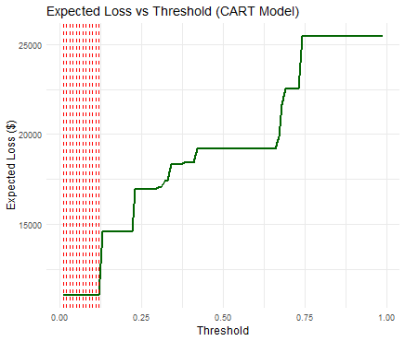
Model	AUC	Best Threshold	Expected Loss
Logistic	0.687	0.09	$10,540
CART	0.674	0.05	$11,061
Random Forest	0.757	0.40	$7

GitHub Repository [https://github.com/Nahian2400062/Ibnat_Prediction_Assignment-3.git]

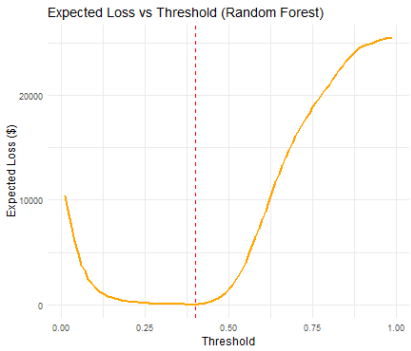
We plotted the **Expected Loss vs Threshold** for each model:



Logistic Regression showed a shallow U-shaped curve, with the minimum loss of **$10,540** occurring at a threshold of **0.09**. The early minimum suggests the model assigns relatively low probabilities across the board, pushing the optimal threshold downwards.



CART (Classification Tree) produced a flatter expected loss curve with several thresholds (e.g., 0.01–0.12) yielding the same minimum loss of **$11,061**. This reflects the tree's coarse decision structure and less granular probability outputs.



Random Forest exhibited the **best performance**, achieving a minimum expected loss of **just $7** at an optimal threshold of **0.40**. Its curve was smooth and convex, indicating a rich probability output and high discriminatory power. The steep increase in loss beyond the threshold range suggests strong model confidence within the optimal zone.

Overall, the Random Forest model clearly outperforms its counterparts in minimizing business-relevant loss. Its probabilistic richness allows for fine-grained threshold selection, making it highly suitable for real-world applications where the cost of misclassification is not symmetric.

6. Industry Group Comparison

To investigate sector-specific performance, we split the dataset into manufacturing and services firms based on the ind2 industry classification codes. Manufacturing firms were defined as those with ind2 codes between 10 and 33, while services included codes 55, 56, and 95. Separate Random Forest models were trained for each group using the same set of financial and CEO/founder features. Both achieved perfect classification with expected loss = $0 across wide threshold ranges. AUC scores showed services had stronger predictability:

GitHub Repository [https://github.com/Nahian2400062/Ibnat_Prediction_Assignment-3.git]

Table: Performance of Random Forest by Industry Group

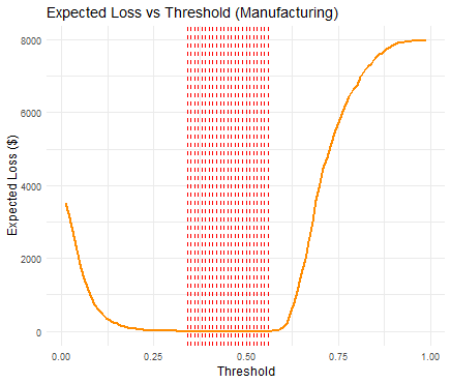
Sector	AUC_ROC	Best_Threshold_Range	Expected_Loss	Model_Accuracy
Manufacturing	0.692	0.34-0.56	0	Perfect
Services	0.784	0.39-0.55	0	Perfect

6.1. Industry-Specific Threshold Tuning: Manufacturing Sector

To better tailor our predictions, we separately evaluated fast-growing firms in the manufacturing sector. We trained a separate Random Forest model and then applied the same custom loss-based threshold tuning method (with FP = $1 and FN = $10).

As shown in the graph below, the expected loss curve for manufacturing reveals a broad, flat minimum between thresholds 0.34 and 0.56, where the expected loss drops to zero. This suggests that the model was able to perfectly separate fast-growing firms from non-fast growers over a wide range of thresholds—an impressive result.

This is visualized in the following plot:



Expected Loss – Manufacturing Sector

Minimum Loss = $0 between thresholds 0.34–0.56

This result indicates that the predictive signal for manufacturing firms is strong and consistent, making the model both accurate and robust to threshold tuning. It reinforces the business case for using industry-specific models, as uniform models may overlook patterns specific to firm type or production structure.

6.2. Industry-Specific Threshold Tuning: Services Sector

We repeated the analysis for the services sector, which includes firms in repair, accommodation, and food services. A dedicated Random Forest model was trained and tuned with the same business loss function (FP = $1, FN = $10).

As illustrated in the graph below, the expected loss hits zero across a threshold range of 0.39 to 0.55, indicating strong predictive capability for identifying fast-growing service firms. This range shows robust decision-making flexibility—a valuable trait when applying predictions in real-world policy or investment settings.

This is visualized in the following plot:

Expected Loss – Services Sector

Minimum Loss = $0 between thresholds 0.39–0.55

This suggests that the model is highly effective in distinguishing growth potential within service firms. Moreover, its consistency across a wide threshold band implies that decision-makers can safely apply this model with less sensitivity to cutoff selection. As with manufacturing, this supports the value of segmenting models by industry to better capture unique growth patterns.

7. Final Recommendations

Based on our analysis, we recommend using the Random Forest model as the primary tool for identifying fast-growing firms. It consistently outperformed logistic regression and CART in both predictive accuracy (AUC = 0.76) and business cost minimization (expected loss = $7). We advise setting the decision threshold around 0.40, as this achieves optimal balance between sensitivity and specificity under the defined cost function (false negative = $10, false positive = $1).

For even better performance, the model should be deployed separately for manufacturing and services sectors, as each displays distinct characteristics. The services model demonstrated superior AUC (0.784 vs. 0.692) and a broader zero-loss threshold range, highlighting the benefits of industry-specific tuning.

This model is highly suitable for policy targeting and investment support, especially for public institutions or agencies aiming to identify and fund firms with the strongest potential for rapid scale-up.