Prediction with Machine Learning for Economists: Assignment 3 Summary Report

Ayazhan Toktargazy, Tatyana Yakushina

1. Summary

This report builds a machine learning model to predict which firms are likely to become high-growth. We use firm-level data from the Bisnode panel (2010–2015) and focus on a 2012 cross-section. The target is defined as at least a 20% increase in turnover between 2012 and 2013, following the OECD definition of high-growth firms.

We test three models: logistic regression (including stepwise and LASSO), Random Forest (RF), and Gradient Boosting Regressor (GBR). Models are evaluated using 5-fold cross-validated RMSE, AUC, and a custom loss function. The loss function puts more weight on false negatives, since missing a high-growth firm is costlier for an investor than backing a firm that doesn't grow.

Among the models, Logit Model M5, which includes 197 variables, achieved the best balance between performance and interpretability, yielding the lowest average expected loss across 5 folds. The performance of the other two models (RF and GBR) was comparable, though slightly lower, likely due to the limited hyperparameter tuning.

2. Target Design & Economic Intuition

We define a firm as "fast-growing" if its turnover (sales revenue) increases by at least 20% between 2012 and 2013, consistent with the official OECD/Eurostat definition of high-growth enterprises (Eurostat Glossary). This choice is grounded in both policy relevance and empirical measurability, and aligns with common international standards.

From a corporate finance perspective, fast growth typically reflects the presence of strong reinvestment capabilities, increasing returns to scale, or strategic expansion opportunities. Identifying these firms early allows investors and policymakers to target resources more efficiently. For venture capitalists or institutional investors, the opportunity cost of missing a high-growth firm often outweighs the loss from investing in a low-growth one, thus motivating our asymmetric loss function in the classification task.

We considered alternative definitions of high growth, including total growth over a two-year horizon (e.g., 2012–2014) and a rolling average of annual growth, calculated as the average of two consecutive one-year growth rates. However, due to data availability constraints and the need to preserve sample size, we adopted a one-year turnover-based definition. More sophisticated indicators (e.g., patents, CEO network data) could improve future models (see Kang et al., 2024), but such data was not available in this project.

3. Data & Feature Engineering

We use the Bisnode panel dataset covering firms from 2010 to 2015. Due to missing or incomplete information in certain years, we focus on the 2012 cross section, which provides the most comprehensive data coverage, with 19,703 firms included across key variables. The target variable is constructed as the growth rate in turnover between 2012 and 2013, consistent with our definition of high-growth firms. Only firms with non-missing turnover data for both years are included in the final sample.

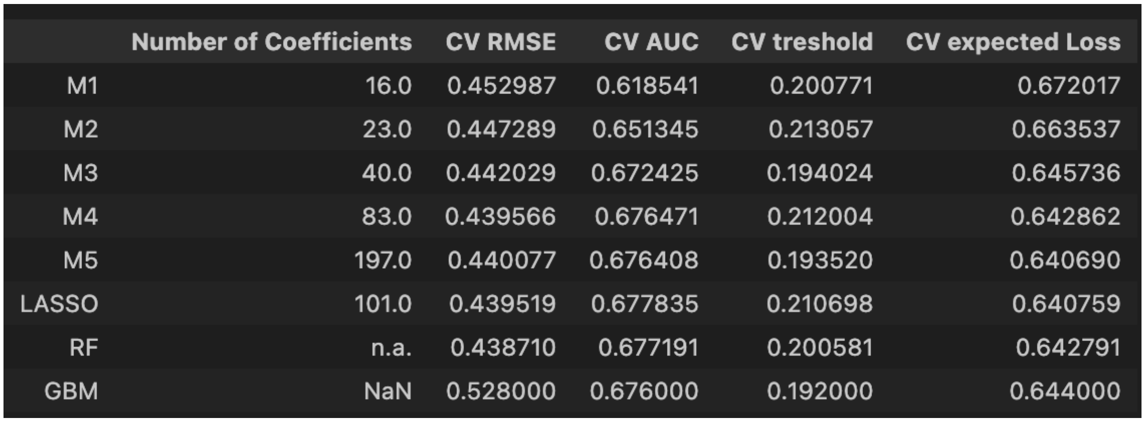
Our feature set draws on multiple categories of firm characteristics, inspired by the firm exit prediction models discussed in class. These include:

- Financial variables: turnover (logged), profit/loss, total and current assets, current liabilities, subscribed capital, shareholder equity, and material expenses.
- Firm characteristics: firm age, legal form, and industry and location fixed effects (region and urban/rural classification), 2-digit industry (NACE) codes.
- Human capital variables: CEO age and gender, number of executives, average wages, and foreign management indicator.
- Quality signals: flags for balance sheet reporting quality and coverage.
- Location variables: region, urban vs. rural classification.

To improve model fit and capture non-linear relationships, we applied log transformations to skewed financial variables, added squared terms for features like firm age and sales, and used one-hot encoding for categorical variables such as industry, region, and firm type. This structured feature engineering approach ensures that both model interpretability and predictive power are preserved, while enabling flexible use across different machine learning methods.

4. Model Comparison

We compare seven models: five logistic regressions of increasing complexity (M1–M5), a LASSO-regularized logistic model, a Random Forest (RF), and a Gradient Boosting Model (GBM). All models are evaluated using 5-fold cross-validated RMSE, AUC, and expected loss based on an investor-focused loss function that places higher weight on false negatives.



	Number of Coefficients	CV RMSE	CV AUC	CV treshold	CV expected Loss
M1	16.0	0.452987	0.618541	0.200771	0.672017
M2	23.0	0.447289	0.651345	0.213057	0.663537
M3	40.0	0.442029	0.672425	0.194024	0.645736
M4	83.0	0.439566	0.676471	0.212004	0.642862
M5	197.0	0.440077	0.676408	0.193520	0.640690
LASSO	101.0	0.439519	0.677835	0.210698	0.640759
RF	n.a.	0.438710	0.677191	0.200581	0.642791
GBM	NaN	0.528000	0.676000	0.192000	0.644000

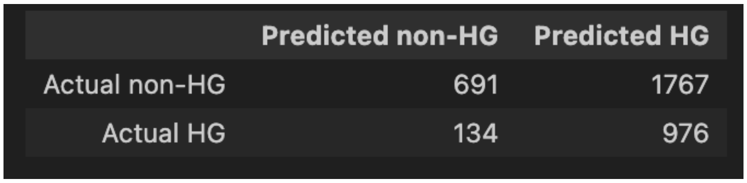
Table 1. Summary of Model Accuracy and Expected Loss

All models performed similarly in terms of AUC. However, the expected loss metric, which reflects the business cost of misclassification, was the deciding factor. Logit M5 offers the best trade-off between interpretability, prediction quality, and business relevance. Ensemble models like Random Forest and Gradient Boosting (GBM) also performed well, but their slightly higher expected loss and greater computational burden made them less favorable for final selection. In particular, GBM's performance was constrained by limited hyperparameter tuning due to hardware limitations; with a more extensive grid search, its accuracy could likely be improved.

Although RMSE was not our primary selection criterion, it provides insight into overall prediction error. GBM showed the highest RMSE among all models, which supports our conclusion that it underperformed slightly in predicting exact probabilities compared to the other approaches.

In the classification step, we applied a custom loss function that reflects an investor's strategic preference: minimizing false negatives (missed high-growth firms) is prioritized over minimizing false positives (investments in non-growing firms). For each model, we selected the optimal classification threshold by minimizing expected loss across five cross-validation folds.

We set the relative cost of a false negative to be four times that of a false positive (FN = 4, FP = 1), reflecting the higher opportunity cost of missing a high-growth firm. Logit M5 achieved the lowest average expected loss under this framework, reinforcing its strength not only in predicting probabilities but also in making final investment-relevant classifications.

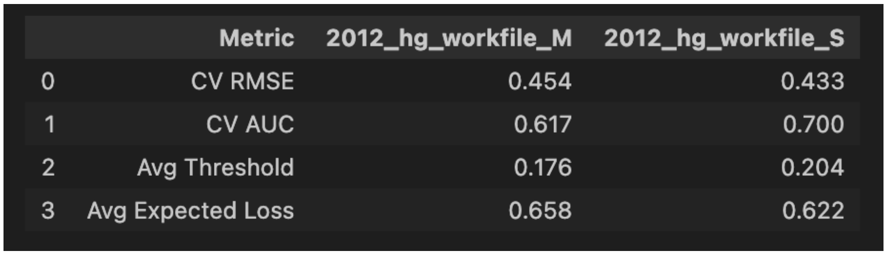


	Predicted non-HG	Predicted HG
Actual non-HG	691	1767
Actual HG	134	976

Table 2. Holdout Set Confusion Matrix – Logit M5

As shown in Table 2, the model correctly identified a large share of high-growth firms while also producing some overclassification of non-high-growth firms. This pattern is aligned with our loss function design: in high-risk-tolerant environments, such as early-stage investment, the downside of missing a high-potential opportunity outweighs the cost of backing a non-scaling firm.

5. Industry-Specific Model Performance: Services vs. Manufacturing



	Metric	2012_hg_workfile_M	2012_hg_workfile_S
0	CV RMSE	0.454	0.433
1	CV AUC	0.617	0.700
2	Avg Threshold	0.176	0.204
3	Avg Expected Loss	0.658	0.622

Table 3. Model Performance Comparison by Industry: Manufacturing vs. Services

To assess model performance across sectors, we applied it separately to manufacturing and services firms, using a consistent loss function and evaluation framework. Random Forest was selected as the final model for its strong predictive accuracy, robustness to overfitting, and relatively low computational demands during tuning.

The model performed better in the services sector (11,842 firms), with higher AUC and lower expected loss, suggesting more predictable growth patterns. This stronger performance may partly reflect the larger sample size, which provides more variation and learning signals for the model. In contrast, the manufacturing sector (5,619 firms) showed weaker performance, possibly due to both its smaller sample size and greater variability in firm size, capital intensity, and production cycles. Allowing for sector-specific loss functions or thresholds could improve classification and better reflect the economic trade-offs in each industry.

6. Conclusion and Future Work

We built models to predict high-growth firms using 2012 data. Logistic regression (M5) achieved the lowest expected loss, while Random Forest performed well across sectors. Model performance, especially for ensemble methods like GBM, was constrained by limited hyperparameter tuning due to computational constraints. Future work could improve accuracy by expanding the tuning grid, incorporating more firm-level indicators (e.g., innovation or CEO networks), enabling dynamic prediction over multiple years, and allowing for industry-specific loss functions to better reflect sector-specific trade-offs.