**Predicting Fast Growth in Firms: A Comprehensive Analysis**

This report presents a detailed study on predicting fast-growing firms using the Bisnode-firms panel data (2010–2015). Our goal was to design a “fast growth” target, engineer features based on corporate finance concepts, build multiple predictive models, and then select the best model according to both probability prediction performance and classification loss. We also demonstrate how the predictive exercise can be split by industry (manufacturing vs. services) to tailor decision-making.

**1. Target Definition and Rationale**

The target variable—fast growth—was defined using a two-year growth rate calculated from sales figures. Specifically, growth was computed as the relative change in sales between a baseline year (2012) and a later year (2014). We set the 75th percentile of the growth rate distribution as the threshold, so that firms with growth rates above this cutoff were labeled as fast-growing.

This approach was chosen for several reasons:

• **Comparability Across Firms:** By using a relative growth metric, we account for the size differences between firms, ensuring that the model captures meaningful changes rather than absolute sales volume.

• **Robustness:** The use of the 75th percentile as a cutoff ensures that the target class is neither too rare nor too common, striking a balance that benefits model training.

• **Alternative Definitions Considered:** Other options included defining growth over a single year or using absolute increases in sales. However, a two-year measure was preferred because it smooths out short-term fluctuations and reflects a more sustained performance trend, which is in line with corporate finance’s emphasis on long-term value creation.

**2. Data Preparation and Feature Engineering**

**Data Cleaning and Sample Design**

Starting with the panel data from 2010 to 2015, the initial steps included:

• Dropping unnecessary variables and filtering out the year 2016.

• Reshaping the data to obtain a panel format with one observation per firm per year.

• Ensuring the sample consisted only of firms with positive sales in 2012 (the baseline year) to compute valid growth rates.

**Creation of the Target and Predictor Variables**

After pivoting the sales data and merging it back into the main dataset, we computed the two-year growth rate. Firms were labeled as fast growing (1) if their growth rate exceeded the 75th percentile and as non-fast growing (0) otherwise.

Key variable transformations included:

• **Log Transformations:** For instance, sales figures were transformed using the natural logarithm (e.g., ln\_sales and sales\_mil\_log) to reduce skewness.

• **Differences and Winsorization:** The change in log sales (d1\_sales\_mil\_log) was winsorized to control for extreme values.

• **Financial Ratios and Flags:** Balance sheet items were scaled by total assets, and flag variables were generated to signal potential data issues (e.g., asset problems) and outliers.

• **Categorical Variables:** Industry classifications and region indicators were converted into dummy variables using techniques such as dummy coding in patsy.

**Engineering Interaction Terms and Higher-Order Features**

Additional matrices were built to capture interactions (e.g., between industry and age or sales) and non-linear effects (e.g., squared terms for log sales). These enriched feature sets were organized into several candidate predictor sets (X1, X2, X3, X4, X5), allowing us to compare simpler models against those incorporating richer interactions and engineered variables.

**3. Model Building and Comparison**

**Overview of Models Developed**

We built a series of models to predict the probability of fast growth:

• **Linear Regression (OLS):** A baseline model was built on the simplest predictor set (X1).

• **Logistic Regression (Logit):** Several logit models were developed (using X1 and X2) to capture the binary nature of the target.

• **LASSO Regression:** Introduced to enforce variable selection in a high-dimensional space.

• **Random Forest (rf\_p):** To capture non-linear relationships and interactions that a linear model might miss.

• **Ensemble Methods:** Gradient Boosting Machines (GBM), XGBoost, and CatBoost were also implemented to optimize prediction accuracy through ensemble learning.

**Cross-Validation and Performance Metrics**

Each model was evaluated using cross-validated performance metrics, including:

• **CV RMSE:** To assess the prediction error in probability estimation.

• **CV AUC:** To measure discriminative ability.

• **CV Classification Threshold and Expected Loss:** A business-specific loss function was defined (with different costs for false positives and false negatives) to compute the optimal classification threshold and the corresponding average expected loss.

The final performance summary table is shown below:

| **Model** | **Number of Predictors** | **CV RMSE** | **CV AUC** | **CV Threshold** | **CV Expected Loss** |
| --- | --- | --- | --- | --- | --- |
| X1 | 11 | 0.4524 | 0.5526 | inf | 0.2906 |
| X2 | 18 | 0.4482 | 0.5983 | 0.5130 | 0.2893 |
| X3 | 35 | 0.4471 | 0.6056 | 0.5296 | 0.2887 |
| X4 | 79 | 0.4452 | 0.6171 | 0.5242 | 0.2861 |
| X5 | 163 | 0.4458 | 0.6184 | 0.5871 | 0.2876 |
| LASSO | 107 | 0.4454 | 0.6207 | 0.5681 | 0.2878 |
| rf\_p | 44 | 0.4443 | 0.6239 | 0.6285 | 0.2850 |
| GBM | 44 | 0.4415 | 0.6406 | 0.5831 | 0.2849 |
| XGBoost | 44 | 0.4415 | 0.6402 | 0.7880 | 0.2873 |
| CatBoost | 44 | 0.4405 | 0.6467 | 0.5436 | 0.2831 |

Among the models, the ensemble methods—particularly CatBoost—yielded the best performance with the lowest CV RMSE and expected loss, and the highest CV AUC. The selection of CatBoost was guided not only by its overall prediction accuracy but also by its stable performance across cross-validation folds and its ability to handle categorical variables efficiently.

**4. Classification Strategy and Business Considerations**

**Loss Function Definition**

Given the business context, misclassifications carry monetary implications. We defined a loss function with specified costs for false positives (FP) and false negatives (FN). The optimal threshold for classification was determined by minimizing the expected loss over five cross-validation folds. This approach allowed us to directly incorporate business risk into the model selection process.

**Performance Evaluation and Confusion Matrix**

For the selected model (e.g., CatBoost), a confusion matrix was computed on a holdout sample to illustrate the trade-off between sensitivity (true positive rate) and specificity (true negative rate). This confusion table, along with ROC curves and loss plots, provides a clear picture of how the model performs in practical settings. Detailed plots (such as ROC curves with optimal thresholds) were generated to aid in visual decision-making.

**Industry Segmentation Analysis**

Recognizing that industry dynamics differ, the analysis was extended to separately evaluate manufacturing and services (including repair, accommodation, and food). The same loss function was applied to both segments to compare model performance. This segmentation helps senior managers understand whether predictive strategies should be tailored to different industry contexts, thereby enhancing the model’s operational relevance.

**5. Discussion and Final Recommendations**

In summary, the analysis demonstrates a systematic approach to predicting fast-growing firms. The detailed data cleaning and feature engineering steps ensured that the models could capture the nuances of firm performance over time. By exploring multiple predictive algorithms—from logistic regression and LASSO to tree-based ensembles—we were able to identify a model (CatBoost) that achieves superior balance between predictive accuracy and expected loss.

The modeling decisions were informed by both statistical performance and business impact. For instance, while simpler models provided a good starting point, the enriched variable sets and interaction terms in the more complex models better captured the underlying non-linear relationships. The classification task, with its emphasis on a customized loss function, directly connects the model outputs to managerial decision-making.

**Final Recommendation:**

For practical deployment, we recommend using the CatBoost model as it exhibits the lowest expected loss (0.2831) and the highest AUC (0.6467) among all models tested. Its performance is robust across different segments, and the ability to adjust the classification threshold based on business loss considerations makes it a valuable tool for identifying high-growth opportunities. Furthermore, the industry-specific analysis suggests that while the overall approach is sound, fine-tuning the model for manufacturing versus services may yield additional benefits.