MPhil in Economics and Data Science

Module: D100/D400 – Fundamentals of Data Science and Research Computing

Candidate Number (BGN): 3365F

Deadline Date: 19th Dec

I confirm that this is entirely my own work and has not previously been submitted for assessment, and I have read and understood the University's and Faculty's definition of Plagiarism (please see links below)

Plagiarism and Academic Misconduct

Plagiarism and correct referencing in dissertations Moodle Econ Update (003).pdf

Actual word count: 1976

Firms’ Credit Ratings Prediction

Introduction

Predicting corporate credit ratings is critical for assessing a company's financial stability and risk. This project uses machine learning models, specifically the Generalized Linear Model (GLM) and LightGBM (LGBM), to automate credit rating predictions based on financial indicators. GLM is a robust statistical model that works well for linear relationships, offering transparency and interpretability. On the other hand, LGBM is a high-performance gradient boosting model that captures complex relationship, especially in handling non-linearities and feature interactions common in financial datasets. The combination of GLM and LGBM allows us to benchmark accuracy, interpretability, and computational efficiency, providing comprehensive insights.

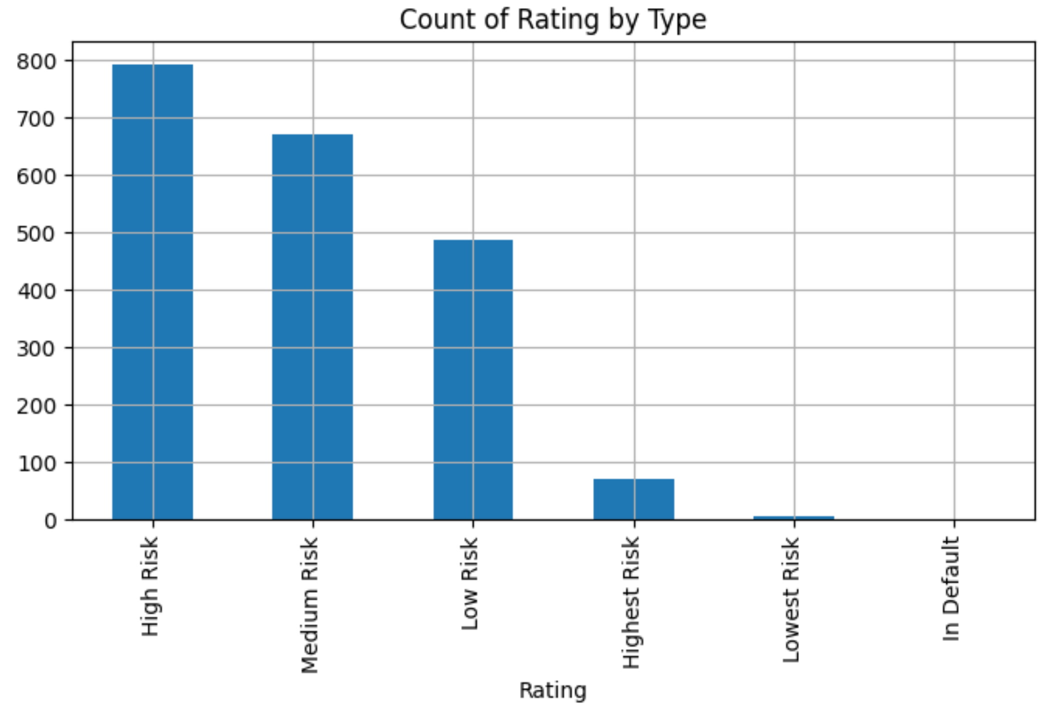
To conclude, LGBM outperforms GLM in accuracy, precision, F1 score, and other metrics. While GLM is interpretable and useful for simpler relationships, LGBM’s ability to handle non-linear patterns and complex data makes it the superior choice for credit rating predictions.

Explanatory Data Analysis and Data Cleaning

The dataset consists of 2029 records with 31 attributes, including: 25 numerical financial indicators and 6 categorical or descriptive features.

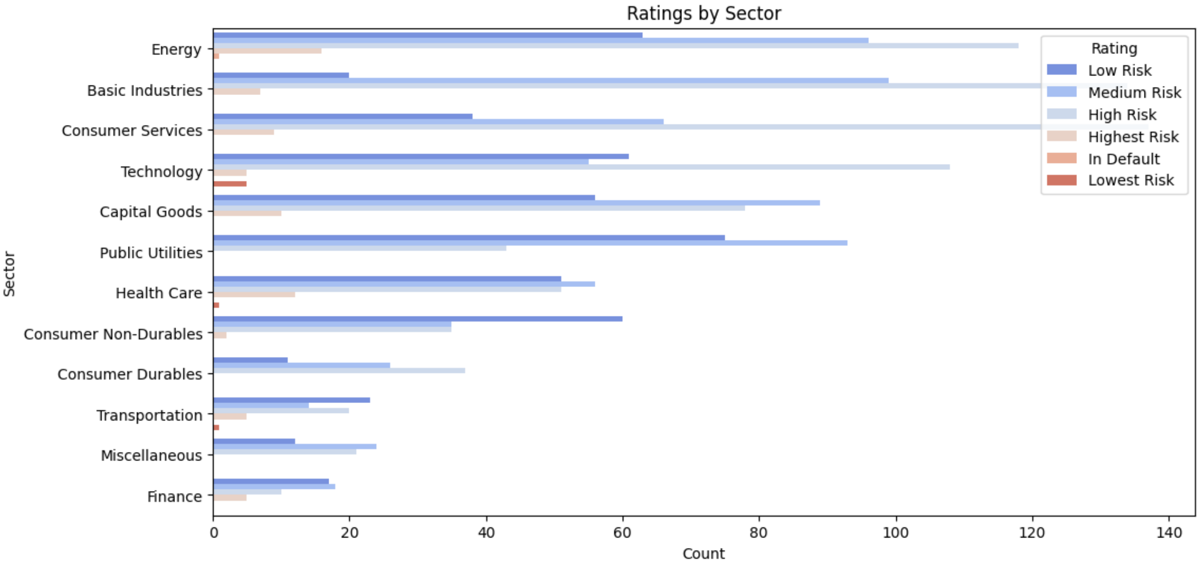
Skewness

Our target variable Rating is a categorical variable, we can firstly observe the distribution of it. The bar chart highlights an imbalance in credit ratings:
1. "Medium Risk" and "High Risk" ratings dominate the dataset.
2. Very few instances belong to "Lowest Risk" and "In Default" categories.



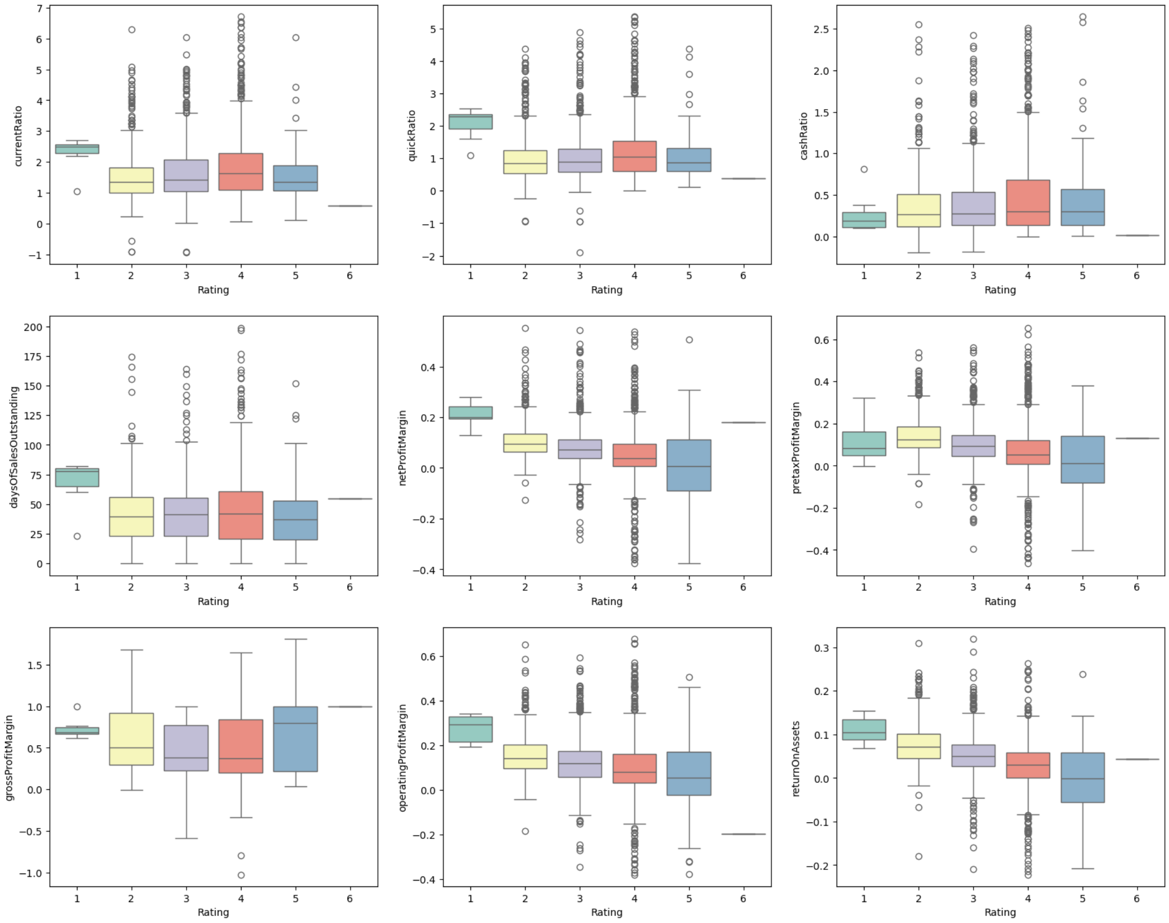
The bellowing chart highlights sector-specific financial stability and risk distribution. Sectors such as Energy and Basic Industries face elevated financial risk, while Public Utilities and parts of Consumer

Services demonstrate more moderate and stable performance. Technology exhibits diversity, with companies spread across all rating levels, including Default. We can use encoding function on this variable since it may be a potential good predictors for credit rating.



Presence of the Outliers

The plots above display a series of boxplots that analyse the distribution of some financial indicators against the Rating variable (higher risk with higher value). Each subplot compares how these financial features behave across the different rating categories.



The central box shows the interquartile range (IQR), representing the middle 50% of the data (from Q1 to Q3).The line inside the box indicates the median (Q2). The whiskers extend from Q1 to the minimum non-outlier and from Q3 to the maximum non-outlier within 1.5 times the IQR. Data points outside the whiskers (beyond 1.5 * IQR) are considered outliers and are shown as individual dots.

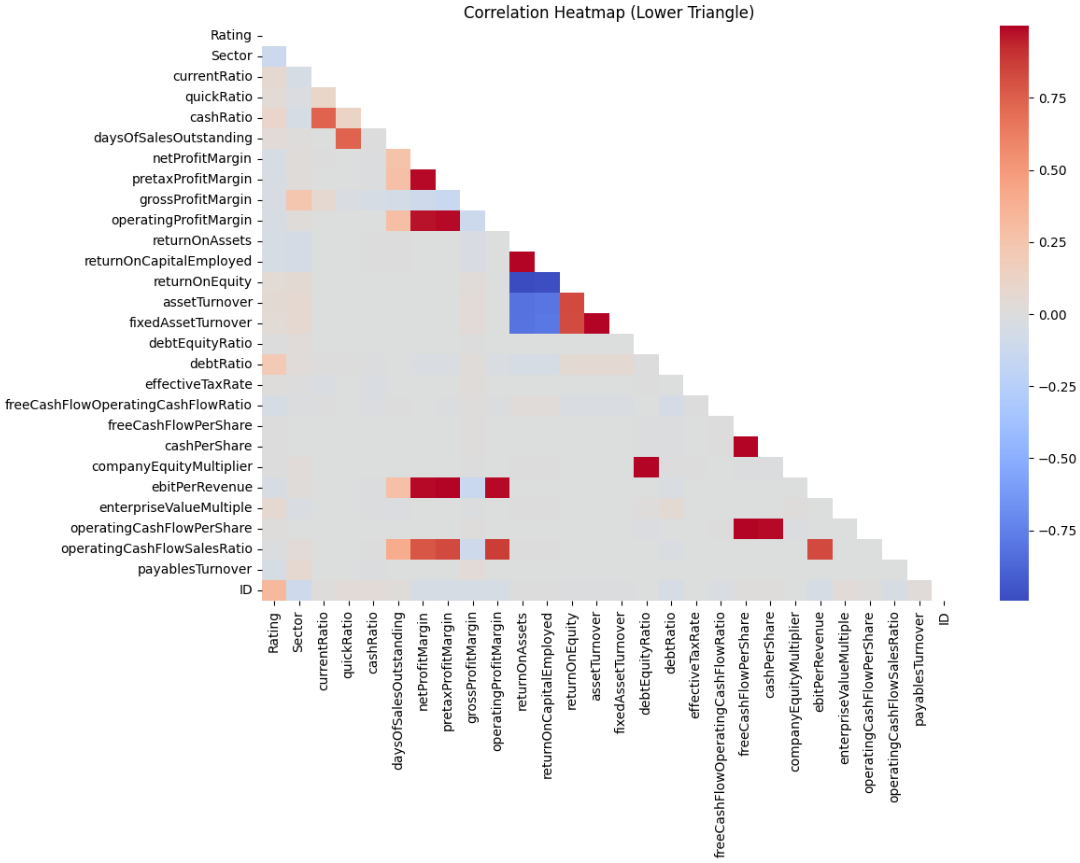
As can be seen, these plots reveal insights into their spread, central tendency, and presence of outliers. Companies with lower credit risk (Lowest Risk and Low Risk) exhibit higher liquidity ratios (currentRatio, quickRatio), positive profitability (netProfitMargin, grossProfitMargin), lower debt levels (debtRatio, debtEquityRatio), and efficient operational performance (returnOnAssets, assetTurnover), reflecting strong financial health. In contrast, high-risk companies (In Default and Highest Risk) show lower liquidity, negative or near-zero profitability, higher debt burdens, and delayed cash recovery cycles (e.g., elevated daysOfSalesOutstanding), with broader spreads and more extreme outliers across financial indicators.

These patterns highlight that liquidity, profitability, and debt management are key differentiators of credit risk, with features like currentRatio, netProfitMargin, debtRatio, and daysOfSalesOutstanding being particularly indicative.

Correlations between each variable

The heatmap visualized the correlation matrix of all numerical features. Notable patterns include:

1. High Positive Correlations: grossProfitMargin and operatingProfitMargin have strong correlations, indicating that these features capture similar profitability insights.
2. Weak Correlations with Target Variable (Rating): Features like operatingCashFlowSalesRatio and companyEquityMultiplier show weak relationships with the target variable, suggesting limited predictive power.



Based on the analysis above, we clean the data as follow:

- Drop the unrelated categorical data: Name, Date, Symbol, Rating Agency Name.
- Transform the 10 types of Rating to 6 Risk levels.
- Assign 6 different numerical values to Rating, the higher the value, the higher Risk.
- Filtered extreme outliers using the 5th and 95th percentiles.
- Encoded the target variable (Rating) and Sector using ordinal encoding.
- Dropped underrepresented classes (Lowest Risk and In Default).
- Removing highly correlated features to reduce redundancy
- Drop missing value

Feature Engineering and Model Training

Feature engineering process

This process combines:

1. Quantile capping and log transformation to handle outliers and skewness in numerical features. The custom transformer class LogCap first determines the lower and upper quantile thresholds using the fit method, where values below the specified lower quantile (e.g., 5%) and above the upper quantile (e.g., 95%) are identified. During transformation, these extreme values are clipped to the quantile thresholds using np.clip to limit the range of the data.
2. After capping, a log transformation with a small constant (0.01) is applied to avoid issues with zero values, compressing large values and reducing skewness while maintaining numerical stability. This process ensures that extreme outliers do not dominate the data distribution and prepares the features for machine learning models that perform better with normalized and symmetric data.

Then the dataset was split into training and validation sets (80/20 ratio).

Model

We trained two models for comparison: GLM and LGBM. GLM (Generalised Linear Model) is a statistical model that extends linear regression by allowing the target variable to have a non-normal distribution and linking it to predictors using a specified link function. It is interpretable and applicable to datasets with linear correlations, making it an important tool for understanding how financial variables influence credit ratings. LGBM (LightGBM) is a high-performance gradient boosting system built for huge datasets with complicated feature relationships. It supports categorical target variables and good at capturing non-linear correlations, making it especially useful for scenarios with complex financial data and patterns.

For GLM, we use Logistic Regression with a multinomial classification approach to handle multiple credit rating classes. The loss function for multinomial Logistic Regression is the multinomial log-loss:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log\left(\hat{y}_{ij}\right)$$

Where N is the sample size, C is the number of classes. y_{ij} is the true label while $\hat{y}_{ij}$ is the predicted probability of sample i belonging to class j. The loss penalizes the model for incorrect class probability predictions, pushing it to maximize the likelihood of the correct class.

For LGBM, we choose the same distribution of the target variable (Multinomial) and the same loss function.

Hyperparameter

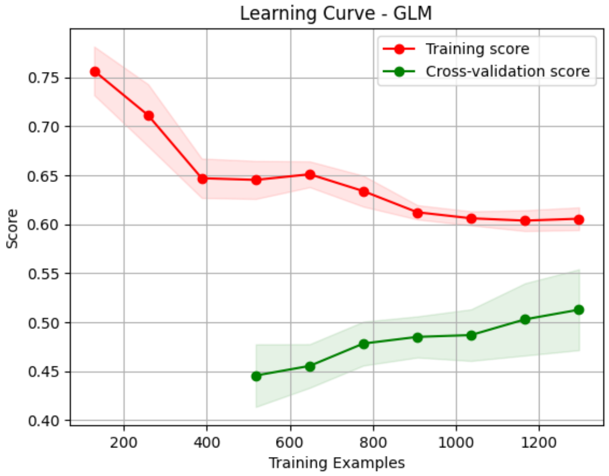
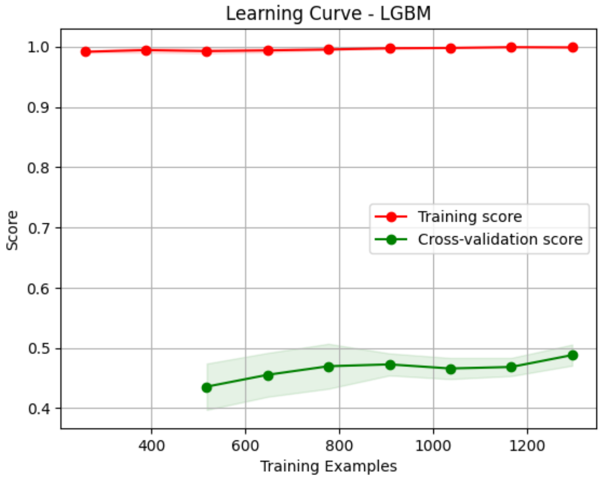
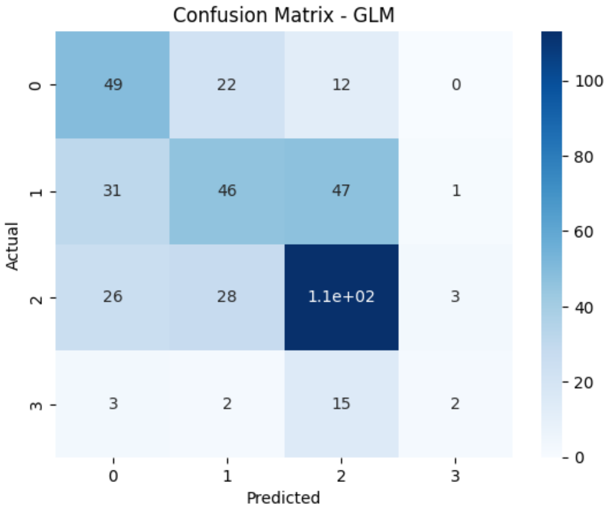
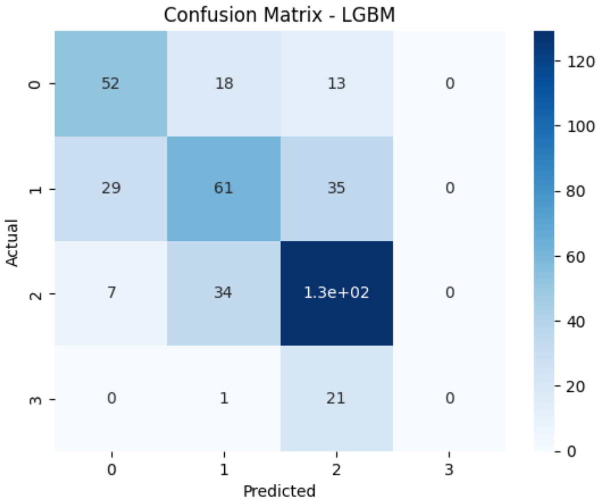
We used GridSearchCV and k-fold cross-validation to optimize hyperparameters for both models. GridSearchCV is a method to systematically tune hyperparameters by exhaustively searching through a predefined grid of values and evaluating each combination using cross-validation to find the best-performing model. K-Fold Cross-Validation is a resampling technique that splits the data into k subsets (folds), where the model is trained on k-1 folds and validated on the remaining fold, rotating this process k times to provide a reliable performance estimate. We explicitly add a term L1-penalty to the optimization problem.

The following parameters were tuned:

GLM	LGBM
classifier__C: Controls the strength of regularization. Smaller values imply stronger regularization.	**classifier__n_estimators:** Number of boosting rounds or decision trees in the ensemble. Fixed at 100 to control training time while maintaining performance.
classifier__solver: Optimization algorithm used for fitting the model. Options: "lbfgs" (quasi-Newton optimization) and "saga" (suitable for large datasets with sparse data).	**classifier__learning_rate:** Shrinks the contribution of each tree by the specified rate. Tuned over [0.01, 0.05] to balance convergence speed and performance.
classifier__max_iter: Maximum number of iterations allowed for the solver to converge. Set to 500 to ensure adequate time for convergence.	**classifier__num_leaves:** Maximum number of leaves in a decision tree. Higher values increase model complexity. Tuned over [15, 31, 45] to optimize model capacity.
classifier__penalty: Type of regularization applied. "l1" penalty encourages sparsity in the model (feature selection).	**classifier__min_child_weight:** Minimum sum of instance weights (or samples) needed in a child leaf. Tuned over [3, 6] to prevent overfitting by limiting the size of leaf nodes.
classifier__l1_ratio: Ratio of L1 regularization in ElasticNet regularization (if applicable). Tuned over [0.1, 0.3] to control the balance between L1 (sparsity) and L2 (ridge) penalties.	**classifier__lambda_l1:** L1 regularization term on weights, encouraging sparsity in the tree model. Tuned over [0.1, 0.3] to reduce overfitting.

Model Evaluation

The performance of both models was evaluated using confusion matrices and learning curves.

<table>
<tr><th>Model</th><th>GLM</th><th>LGBM</th></tr>
<tr><td>Learning Curve</td><td>
</td><td>
</td></tr>
<tr><td>Actual vs Predicted</td><td>
</td><td>
</td></tr>
<tr><td>Performance</td><td>Accuracy: 0.53
Precision: 0.52
F1_score: 0.51
Cohen_kappa: 0.29
Mcc: 0.29</td><td>Accuracy: 0.61
Precision: 0.57
F1_score: 0.58
Cohen_kappa: 0.40
Mcc: 0.40</td></tr>
</table>

Class	Credit Rating
0	Low Risk
1	Medium Risk
2	High Risk
3	Highest Risk

Learning Curve

For GLM, the learning curve shows that the training score decreases as more data is introduced, stabilizing around 60%. The cross-validation score starts low (~45%) but gradually improves, indicating underfitting and limited generalization capability. The gap between the training and validation scores suggests that GLM struggles to capture complex relationships in the data.

For LGBM, the training score remains consistently high (near 100%), showing that LGBM learns the training data well. However, the cross-validation score stabilizes at approximately 50%, indicating better generalization compared to GLM but still room for improvement. The smaller gap between the training and validation scores suggests LGBM may slightly overfit the training data.

Confusion Matrix

For GLM, the confusion matrix reveals significant misclassifications across all classes, particularly in Class 1 and Class 2. Misclassifications are evenly spread across incorrect classes, showing GLM's limited ability to differentiate between similar ratings.

For LGBM, it shows improved classification, particularly in Class 2 (predicted correctly 130 times). Misclassifications are reduced for Class 1 and Class 0, but there are still slight errors in Class 3 predictions (but it's probably due to the skewness of the data). The performance for smaller classes is notably better in LGBM compared to GLM.

- Accuracy: LGBM achieves higher accuracy (0.61) compared to GLM (0.53).
- F1 Score: LGBM demonstrates better balance between precision and recall, achieving an F1 score of 0.58 versus 0.51 for GLM.
- Cohen's Kappa & MCC: Both metrics (0.40) confirm that LGBM performs significantly better in aligning predictions with actual values.

To conclude, LGBM generalizes better on the validation data compared to GLM, as seen from the higher validation scores and reduced misclassifications. LGBM improves accuracy for majority classes, particularly Class 2, while GLM suffers from high misclassification rates. GLM is simpler and interpretable but struggles with complex patterns in the data. LGBM, being a gradient boosting model, captures non-linear relationships and performs better overall.

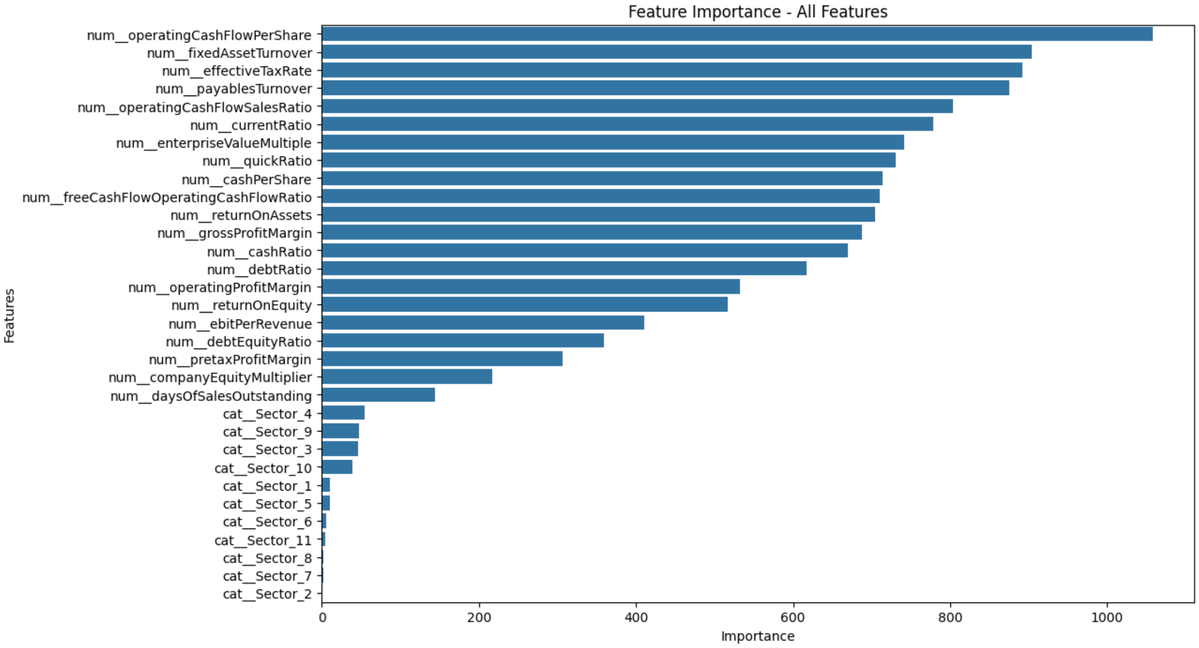
Feature Selection and Interpretation

Partial Dependence Plots (PDPs) were used to interpret the effect of key features in the LGBM model.

Feature Importance

The feature importance plot reveals the most influential variables contributing to the LGBM model's credit rating predictions. Among all features, num_operatingCashFlowPerShare ranks as the most important, indicating that a company's operating cash flow relative to its share value has a significant positive impact on credit ratings. Features like num_fixedAssetTurnover and num_effectiveTaxRate follow closely, reflecting the importance of efficient asset utilization and effective tax management. Additionally, num_payablesTurnover and num_operatingCashFlowSalesRatio emerge as other top contributors, emphasizing the role of liquidity and cash flow efficiency in creditworthiness.

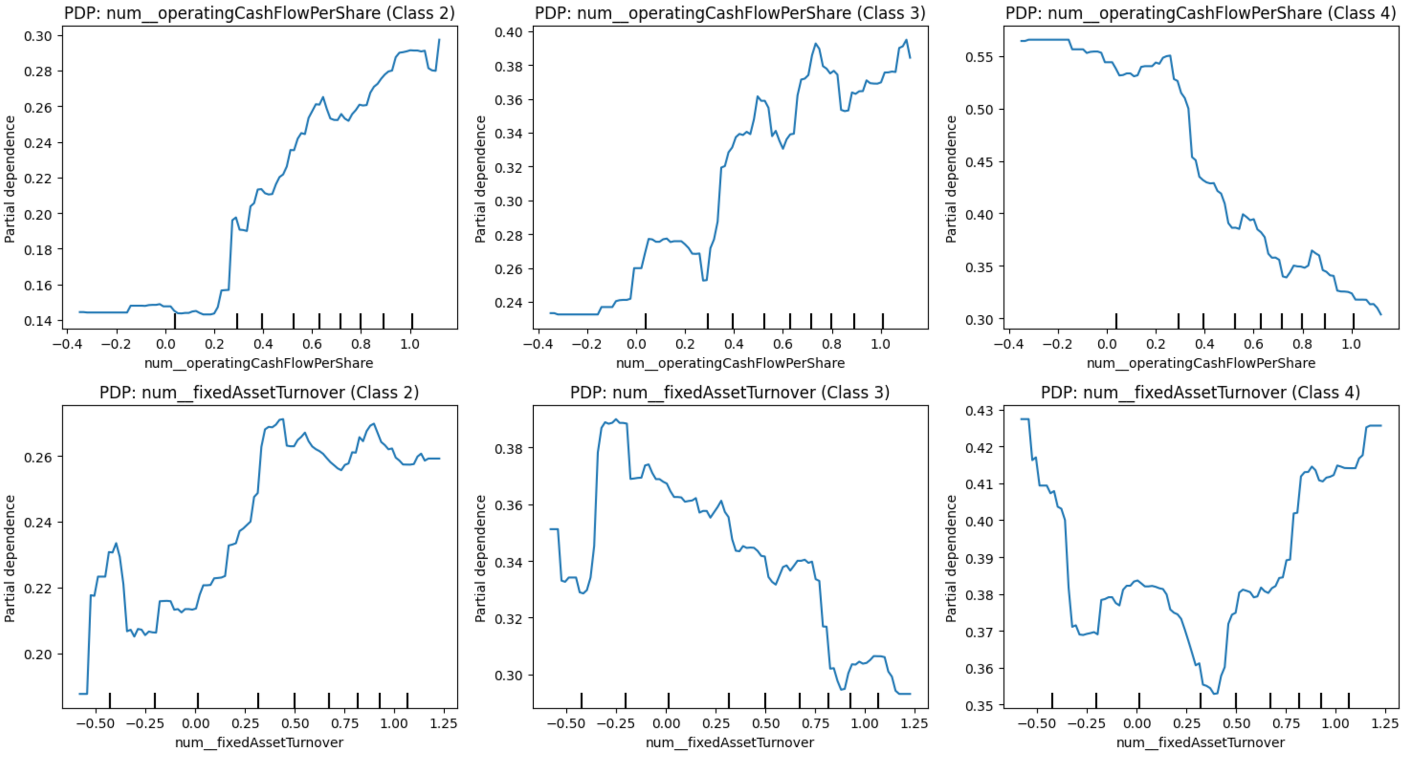
Interestingly, categorical features, such as cat_Sector, have relatively low importance compared to numerical features. This suggests that while a company's sector provides some contextual information, financial ratios and operational metrics are far more critical in determining credit ratings.

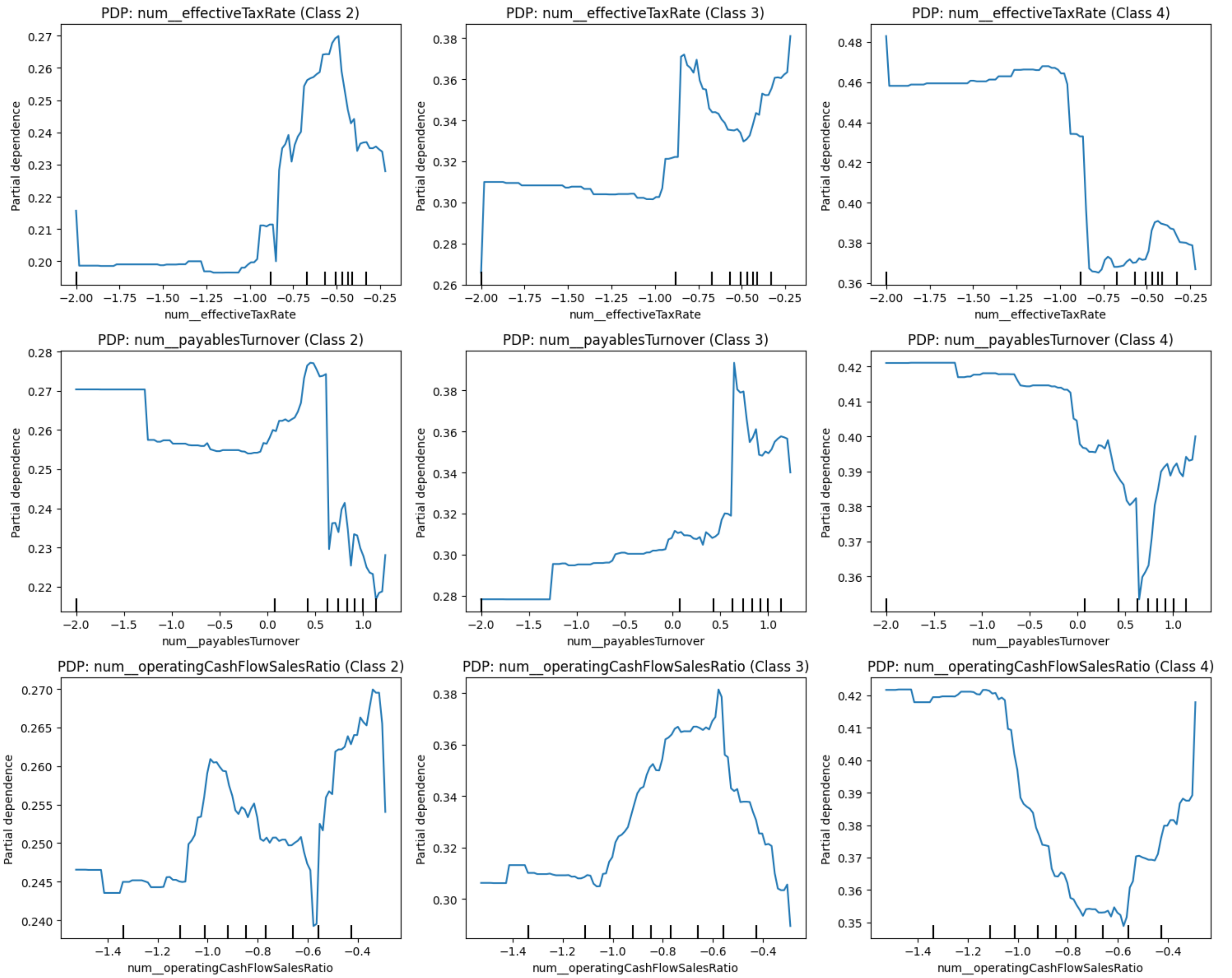


However, we tried to eliminate the Sector from the predictors, the performance of the model does not increase. It may be useful to look at the Partial dependence plots for further investigation.

Partial Dependence Plot (PDP) Analysis

The Partial Dependence Plots (PDPs) provide deeper insights into how the most important features influence predictions across different credit rating classes. The increasing trend represents the variable is positively related with the likelihood of a firm being classified into relatively class of rating, vice versa. To conclude, the findings highlight that financial efficiency metrics, such as cash flow ratios, asset turnover, and tax rates, are the primary drivers of credit ratings. These features exhibit non-linear relationships, where specific thresholds significantly influence predictions, especially in Class 3 and Class 4 ratings. Notably, features related to cash flow (e.g., operatingCashFlowPerShare) consistently show strong positive impacts on lower-risk ratings while reducing the likelihood of high-risk classification





Discussion and Future Improvements

Overall, the LGBM model's ability to capture these complex patterns provides a significant advantage over simpler models like GLM, emphasizing the value of advanced machine learning techniques in credit rating prediction.

One limitation is that I ignore the time invariant to maintain the amount of training data. Some firms may appear several times in the dataset with different time point and this may potentially cause some biases for seeking the estimators. Further study could consider using more advanced machine learning model for this panel data. Additionally, incorporating macroeconomic indicators, such as interest rates and GDP growth, would provide a more holistic understanding of factors influencing credit ratings. Addressing class imbalance through techniques like SMOTE can ensure fair representation of all rating categories, improving predictive accuracy. Finally, experimenting with other methods, such as XGBoost and Random Forest, would allow for performance benchmarking and validation of results.