LightGBM approach to Enhancing Lending Decisions

Nate Adams / QTM 347

Introduction

- **Problem Statement:** Enhancing the accuracy of mortgage approval predictions for Georgia homeowners
- **Project Goal:** Utilize advanced machine learning techniques to reduce financial risks and support fair lending practices.
- **Significance:** Accurate predictions help financial institutions mitigate risks and ensure loans are offered to credible applicants, thus promoting financial stability.
- **Scope:** Analysis involves 501,310 records originally, narrowed down to 399,658 instances after data cleaning, across 25 key features.
- **Motivation:** Address the high stakes involved in mortgage lending by improving prediction reliability, which benefits both lenders and applicants.

Data Overview

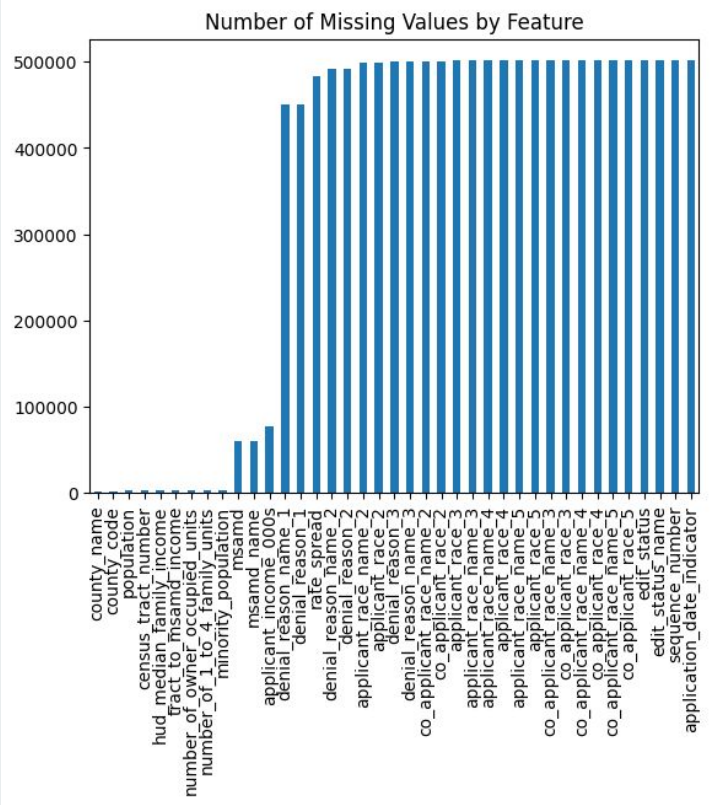
Mortgage Data (HDMA): all records (applications, denials, originations, institution purchases) for Ga homeowners in 2017

Original Dataset: 501310 instances, 78 features

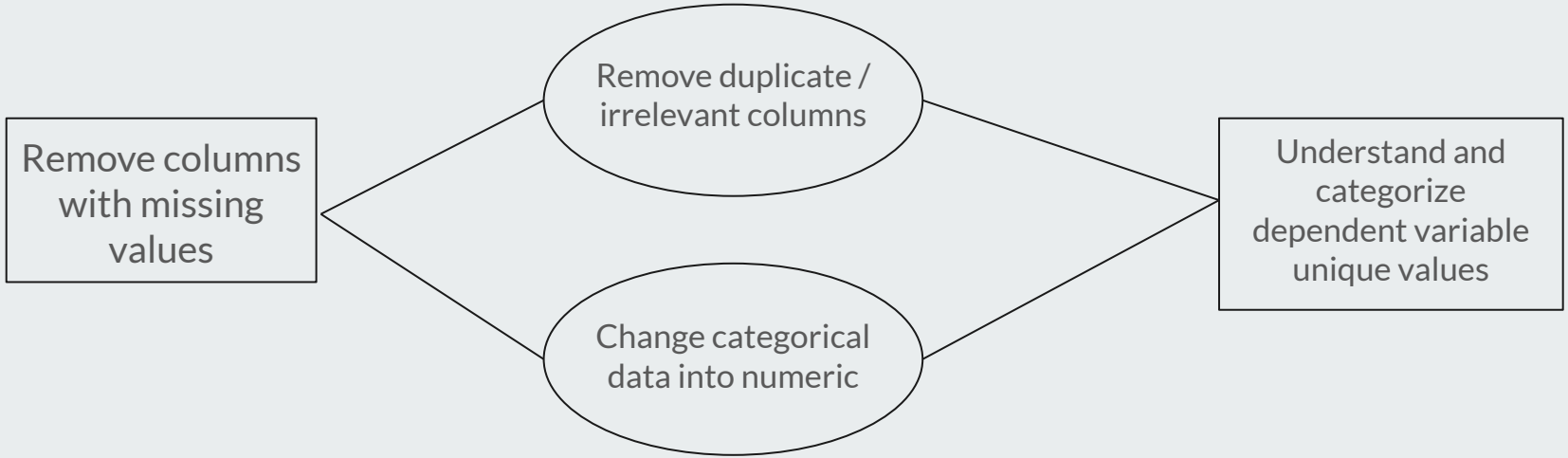
Pared Dataset: 399658 instances, 25 features

Duplicates: Many columns have a corresponding _name column which details what the adjacent column numbers mean (covariance of 1)

Missing Values: 34.79% of data points are blank in the original dataset, with multiple columns having a very high percentage of NaN

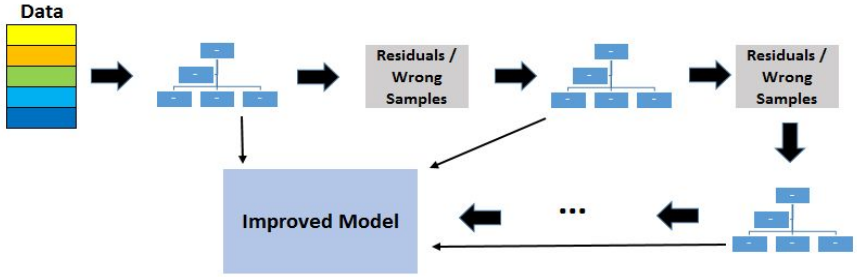


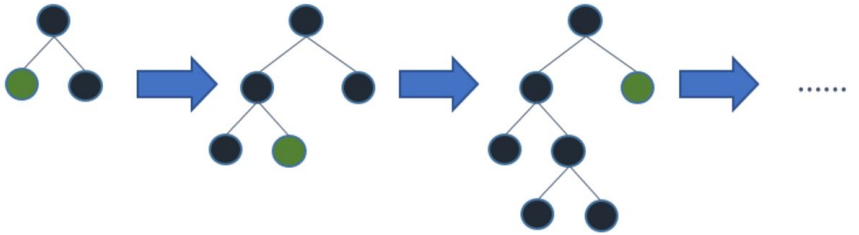
Preprocessing and Data Cleaning



LightGBM: Overview

- **Definition:** LightGBM stands for Light Gradient Boosting Machine, a fast, distributed, high-performance gradient boosting framework based on decision tree algorithms.
- **Advantages Over Other Models:**
  - **Efficiency:** Handles large-scale data with lower memory consumption than XGBoost.
  - **Speed:** Faster training speed and higher efficiency with gradient-based one-sided sampling and exclusive feature bundling.
  - **Accuracy:** Produces more complex models with an improved accuracy due to leaf-wise growth strategies..
- **Popularity:** Widely adopted for its performance, scalability, and efficiency

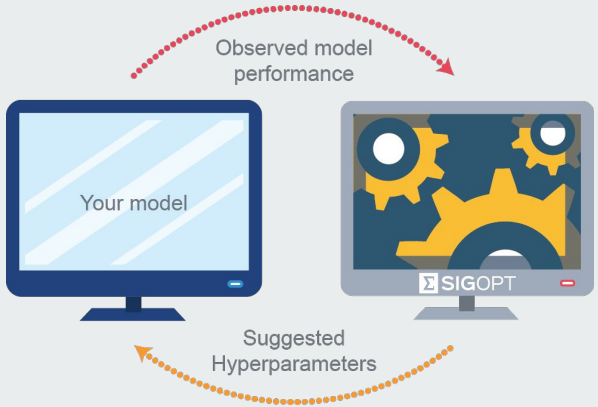




Leaf-wise tree growth

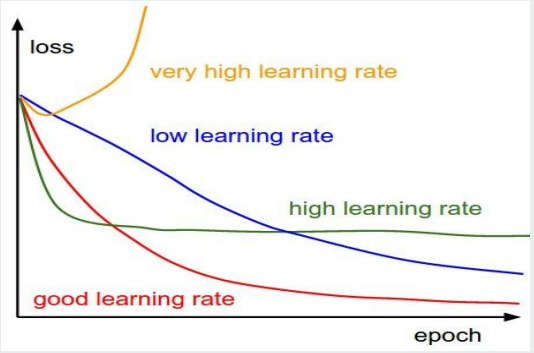
Hyperparameter tuning

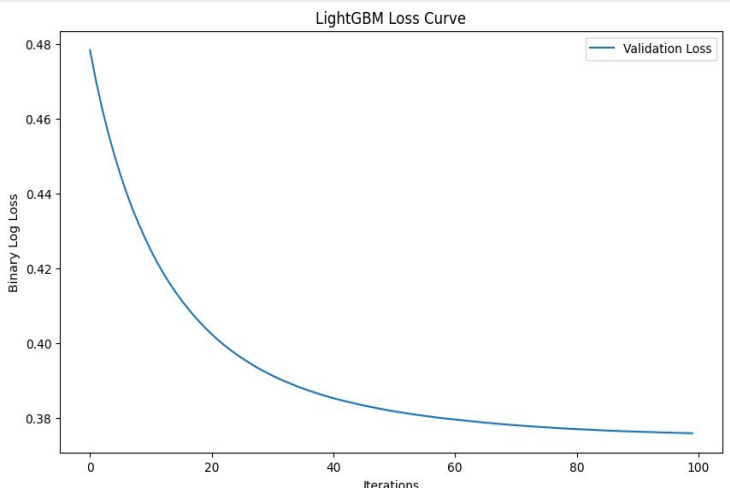
- **Hyperparameters Configuration:**
  - *num_leaves:* Set to 30 to balance model complexity and training efficiency, aimed at capturing subtle patterns without overfitting.
  - *learning_rate:* Fixed at 0.07 to ensure a moderate pace of learning, minimizing the risk of overshooting the minimum loss.
  - *n_estimators:* 100 trees chosen to provide ample learning opportunity while avoiding excessive computation and overfitting.
- **Rationale for Settings:** These parameters were selected after initial experimentation indicated that minor adjustments within reasonable ranges did not yield significant performance changes, suggesting an optimal balance.



Model Performance

- **Performance Metrics:**
  - *Accuracy:* Achieved an accuracy of 83%, moderately better than baseline assumption that everyone gets a mortgage (80%)
  - *ROC AUC:* Scored 0.82, indicating good discriminatory ability between the classes across various threshold settings.
- **Stability of Results:** Consistent performance across different validation folds confirms the model's stability and reliability.
- **Impact of Hyperparameters:** Discuss how the chosen hyperparameters helped in achieving a balance between learning efficiently and preventing overfitting.
- **Comparison with Baseline:**
  - The model's performance exceeds the simplistic baseline assumption that all applicants are approved (unlike a simple regression), which underscores the effectiveness of the LightGBM approach when it comes to messy data
  - Emphasizes the importance of ML in enhancing decision-making over more naive approaches, especially in risk-sensitive sectors like mortgage lending where margins are everything



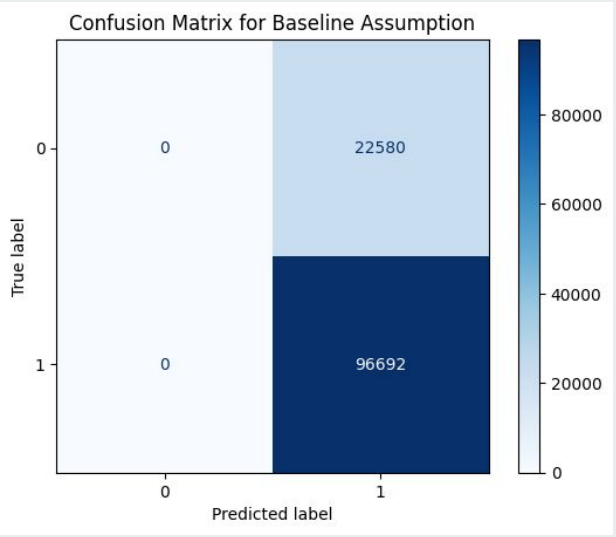


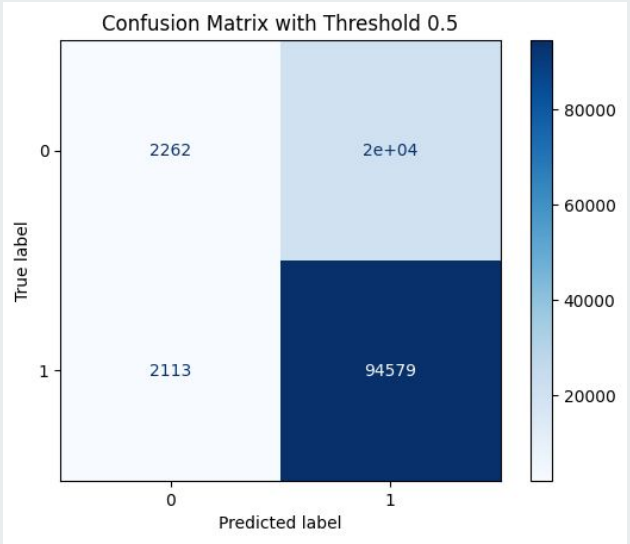
Class Imbalance and Threshold

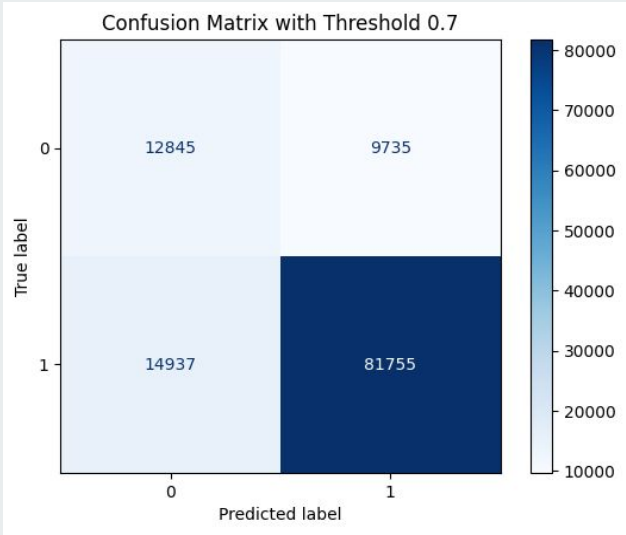
- **Initial Observation:** The dataset exhibited a significant class imbalance, with a much higher number of positive instances (loan approvals) compared to negative instances (loan denials).

- **Adjusting the Decision Threshold:**
  - **Need for Adjustment:** Despite class weighting, the model exhibited a tendency to produce more false positives than desirable, likely due to the inherent challenges of dealing with imbalanced data.
  - **Threshold:** Increased the decision threshold for predicting positive classes (loan approvals) from the default 0.5 to 0.6. This adjustment means that a loan application now needs a higher probability score to be classified as approved.
  - **Impact on Model Performance:**
    - **Reduction in False Positives:** Raising the threshold significantly reduced the number of false positives, aligning more closely with risk management objectives.
    - **Trade-off Considerations:** While the increase in threshold decreased false positives, it was necessary to monitor the effect on false negatives (denying loans that should have been approved), ensuring that the balance achieved maximized overall decision-making accuracy.

Confusion Matrices





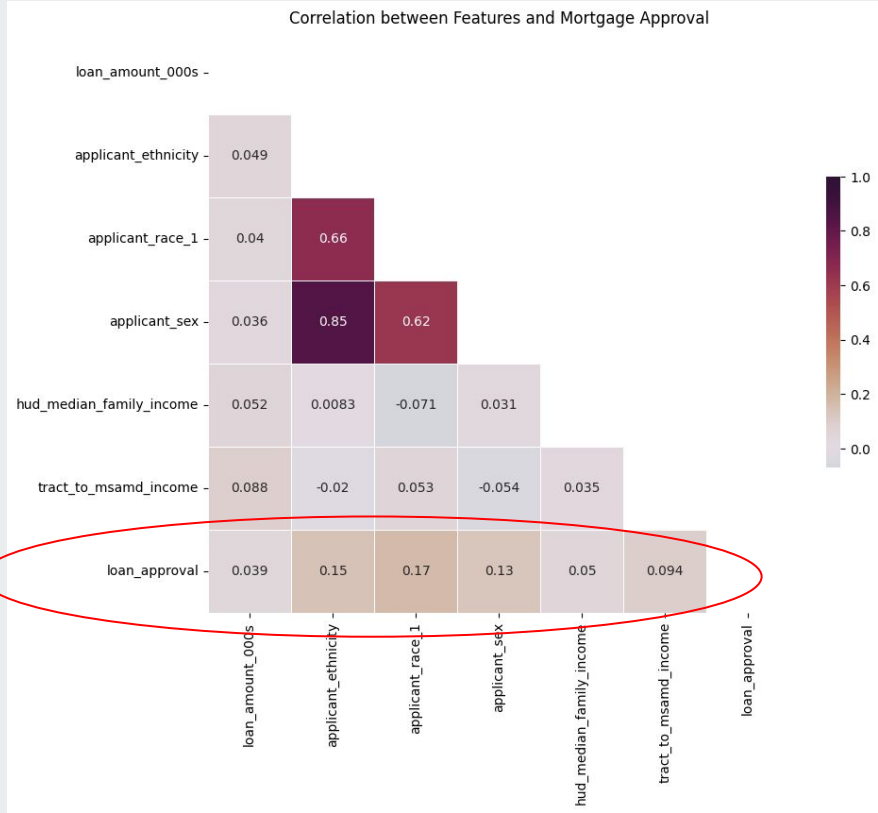


Challenges and Limitations

- **Data Quality and Completeness:**
  - Encountered issues with missing data and inaccuracies in the dataset, which required extensive preprocessing and could potentially affect the model's predictive accuracy.
  - Limitations in data collection methods may lead to biases in the dataset, influencing the model's performance and generalizability.
  - Low correlation with target variable for many of the features made it increasingly difficult to predict positive or negative outcome, as well as lack of key features such as credit score

- **Model Complexity and Interpretability:**
  - The complexity and flexibility of LightGBM, while beneficial for handling diverse data types and large datasets, poses challenges in model interpretability, making it difficult to fully understand decision-making processes and feature influences.

- **Class Imbalance:**
  - Despite efforts to address class imbalance through techniques like class weighting and threshold adjustment, achieving a perfect balance is challenging, and trade-offs between different types of errors (false positives and false negatives) remain a concern.



Significance and Conclusion

- **Enhanced Decision-Making:**
  - The project demonstrates the potential of advanced machine learning techniques like LightGBM to improve mortgage approval predictions, providing lenders with a powerful tool to make more informed and precise lending decisions.
- **Risk Reduction:**
  - By reducing false positives (especially through higher threshold), the model helps mitigate financial risks associated with bad loans, contributing to the overall stability of the financial institutions and protecting them from potential defaults.
- **Broader Impact:**
  - Beyond financial applications, the methodologies and insights from this project can be applied to other areas requiring risk assessment and decision analytics, showcasing the versatility and broad applicability of machine learning in business and economics.

References

https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

https://datascience.stackexchange.com/questions/54043/differences-between-class-weight-and-scale-pos-weight-in-lightgbm

https://www.evidentlyai.com/classification-metrics/classification-threshold#:~:text=The%20classification%20threshold%20in%20machine,not%20assign%20the%20label%20directly.

Dr. Xiong