MGMT 571 DATA MINING FINAL PROJECT

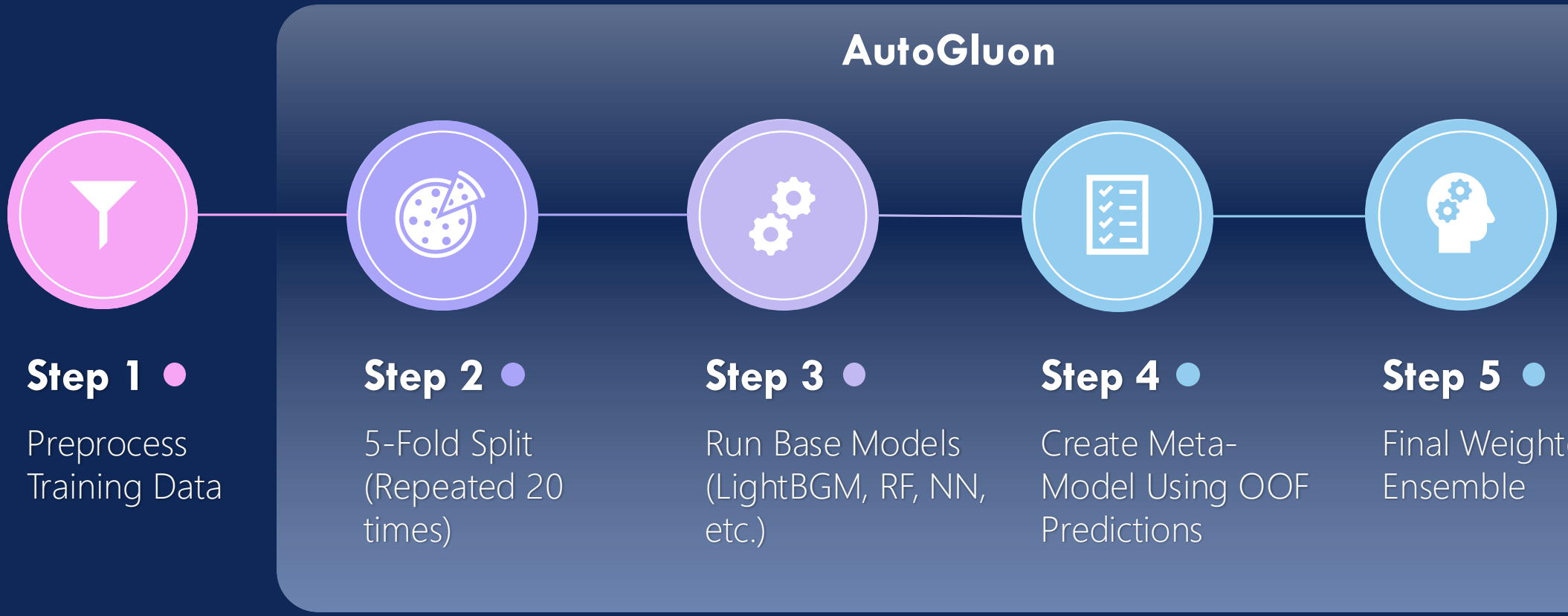
Neural Nomads

Satish | Sana Majeed | Iscel Manalo

CONTENTS

Final Score
0.9231

MODEL SUMMARY



DATA PREPROCESSING

Handling Infinite Values

To avoid errors during model training, all positive and negative **infinite values** in both the training and test data **were replaced with NaN** (missing values)

Selecting Financial Attributes

By explicitly defining the attribute list, the **model focuses on the relevant financial indicators** and ignores any non-predictive columns such as IDs or other metadata that might be present

Reducing Skewness with the *arcsinh* Transformation

To reduce skewness and stabilize variance, the **inverse hyperbolic sine (arcsinh) transformation** was applied

Several preprocessing steps were applied to clean and transform the data to ensure that the model can learn effectively and that issues such as extreme values and skewness are handled appropriately.

WHY AUTOGLUON?

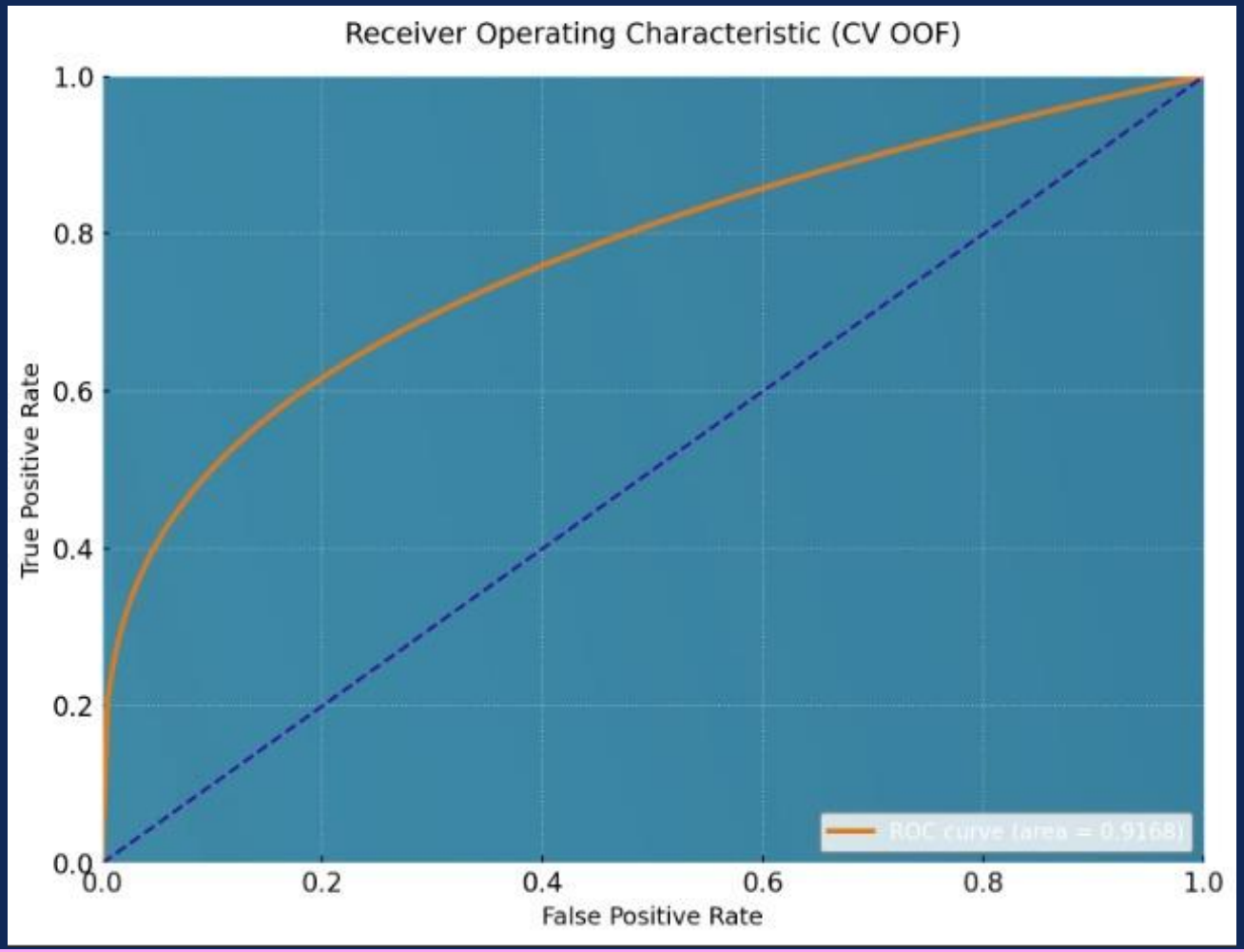
AutoGluon automates the running of different algorithms and hyperparameters.

It can train multiple models such as gradient-boosted trees (LightGBM, CatBoost, XGBoost), random forests, neural networks, and others, and then intelligently ensemble the best ones.

More information: https://github.com/autogluon/autogluon

AutoGluon typically leads to strong performance with much less manual effort.

VALIDATION



The high ROC AUC value indicates that the model is effective at ranking bankrupt firms above non-bankrupt firms in terms of predicted risk.

PROJECT INSIGHTS

- Financial-risk data behaves in a very non-linear way, meaning there isn't just one obvious pattern pointing toward bankruptcy.
- Tuning the model with extended time is beneficial accuracy.
- Simpler models were fast but missed important signals, while very complex models sometimes overfit unless controlled carefully.
- Letting AutoGluon automate model selection, tuning, and ensembling saved a huge amount of manual effort and ultimately produced a model that was both accurate and dependable.
- Datasets with many correlated variables and subtle patterns an automated multi-model approach performs far better than choosing one algorithm by intuition alone.

THANK YOU