ETC3250 Project
The 6ers

Huize Zhang, Monika Mohenska,
Mitchell Ryan Ong-Thomson, Rob Oates,
Michael Ciaravolo, Lindsay Robinson

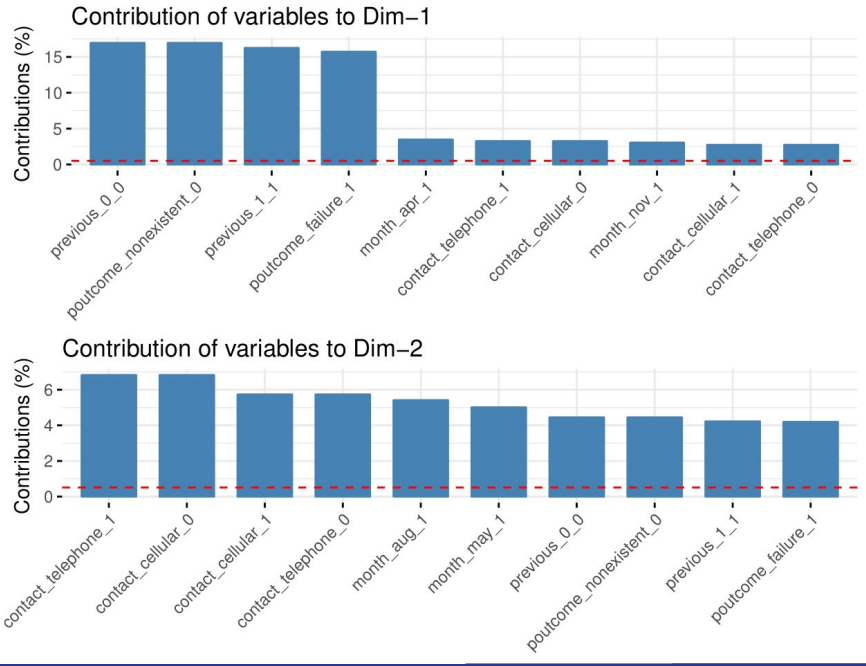
The Aim and The Dataset

Imbalanced Response Variable

High Dimensionality

Sparsity

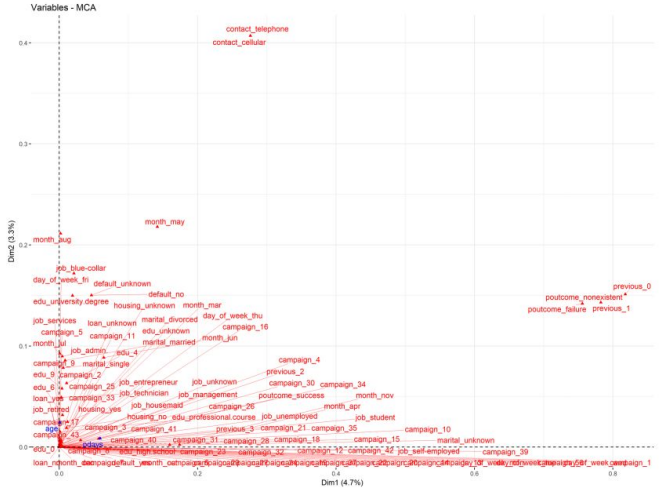
Difficult Regressors

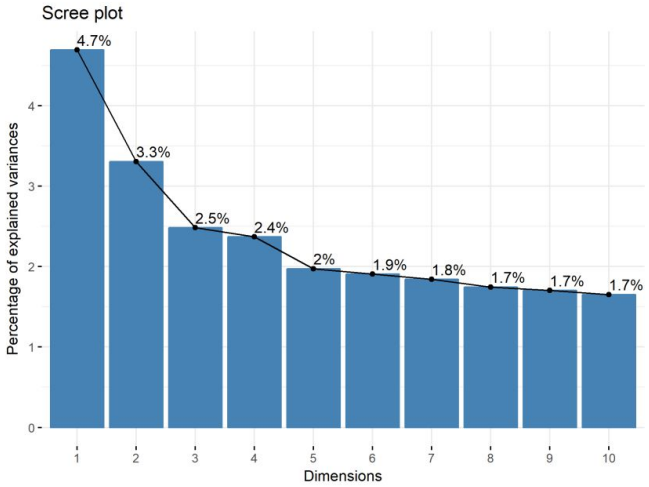


Dimensionality Reduction

PCA Essentially Unavailable

Exploration into MCA





Attempted Models - Logit & LDA/QDA

Logistic Regression

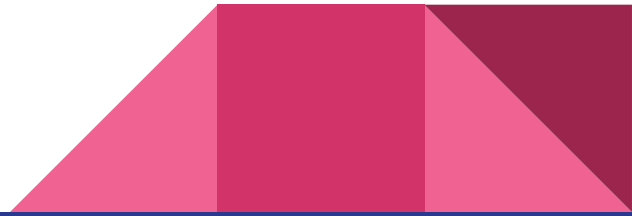
- Non-linear
- Additive Model
- Non-Monotonic

Superior performance to other linear models.

LDA and QDA

Bayesian Predictor Space Separation Methods

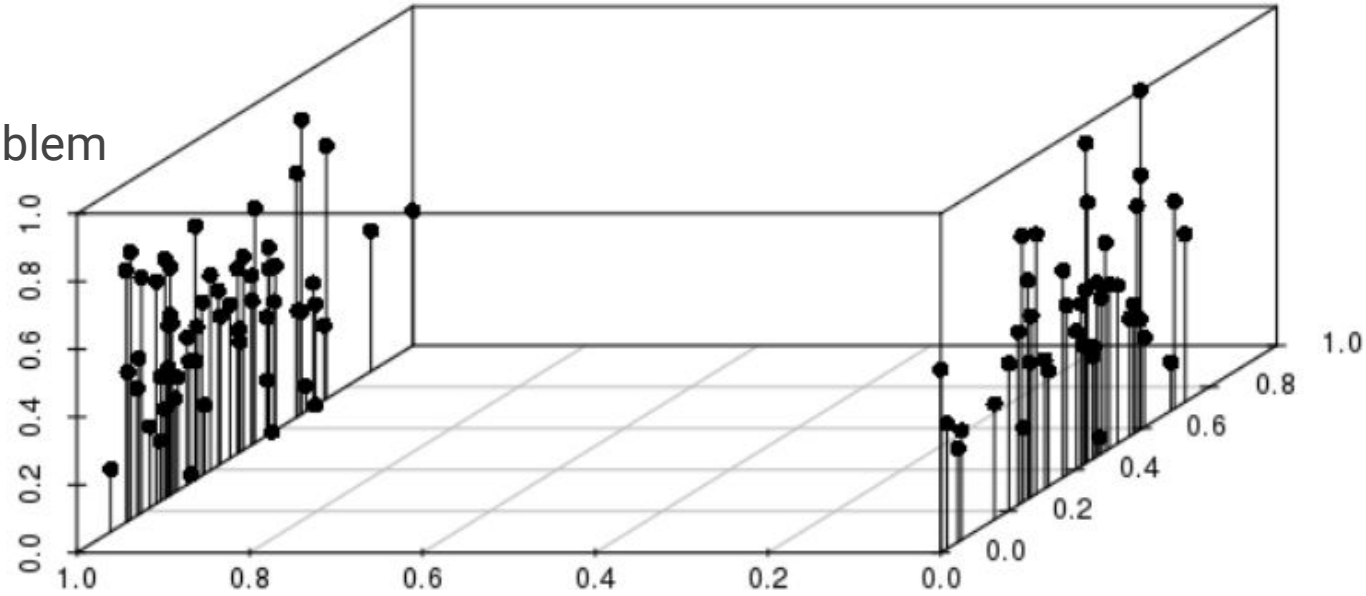
- Computation of a Single Centroid
- Continuous Decision Boundary



Attempted Models - K Nearest Neighbours

Binary Dimension Problem

Complex Decision Boundaries



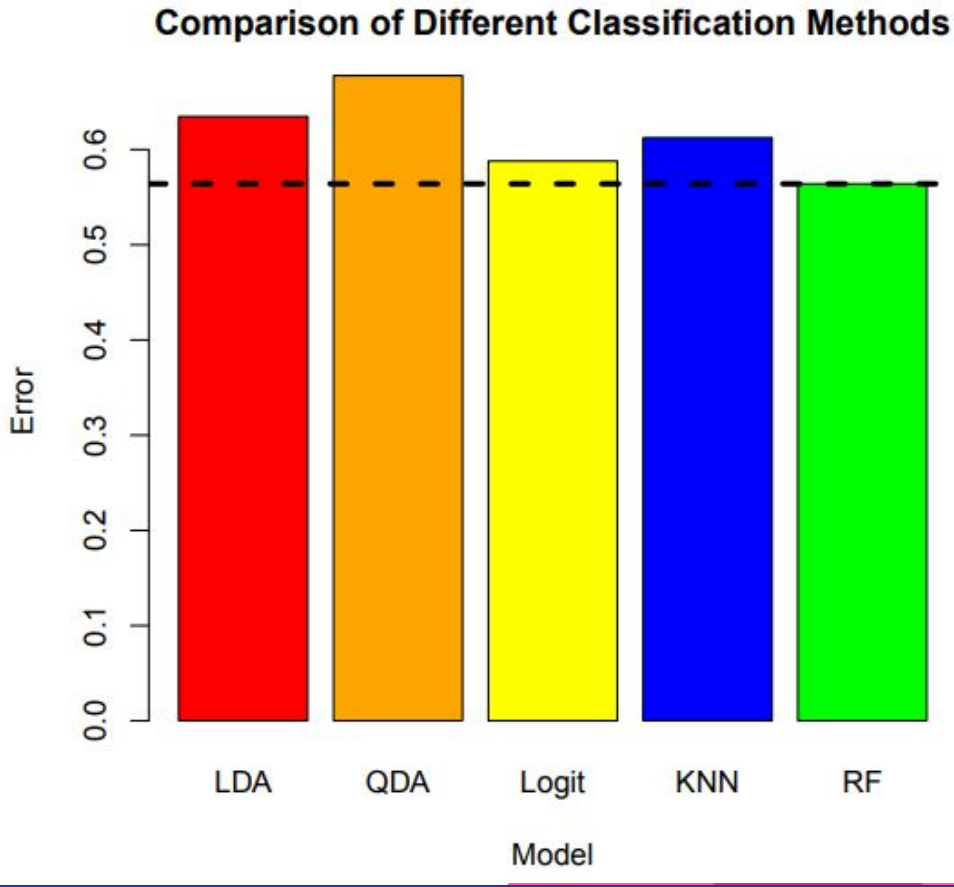
Majority Vote Problem

Random Forest

High Dimensionality

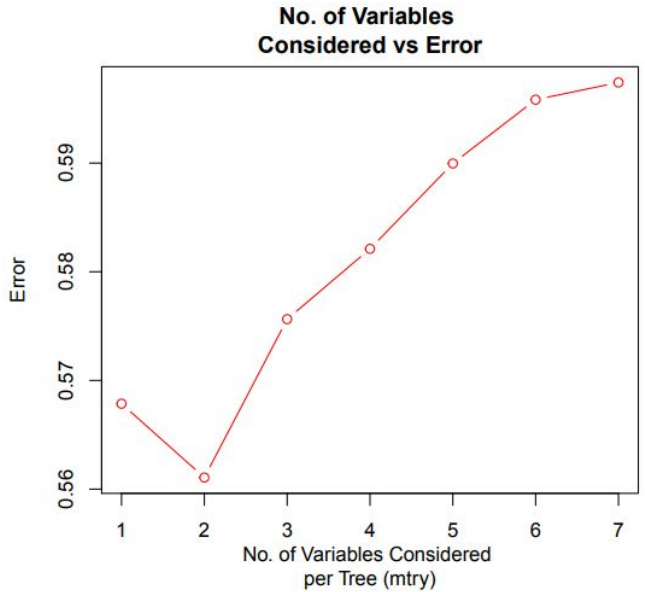
Imbalanced Dataset

Sparsity

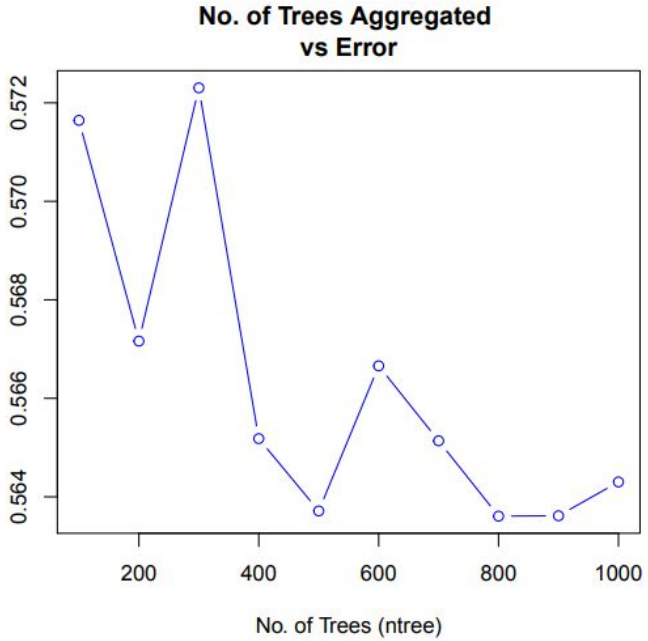


Optimisation of Random Forest

Number of Variables Consider per Tree



Number of Trees Aggregated in Random Forest

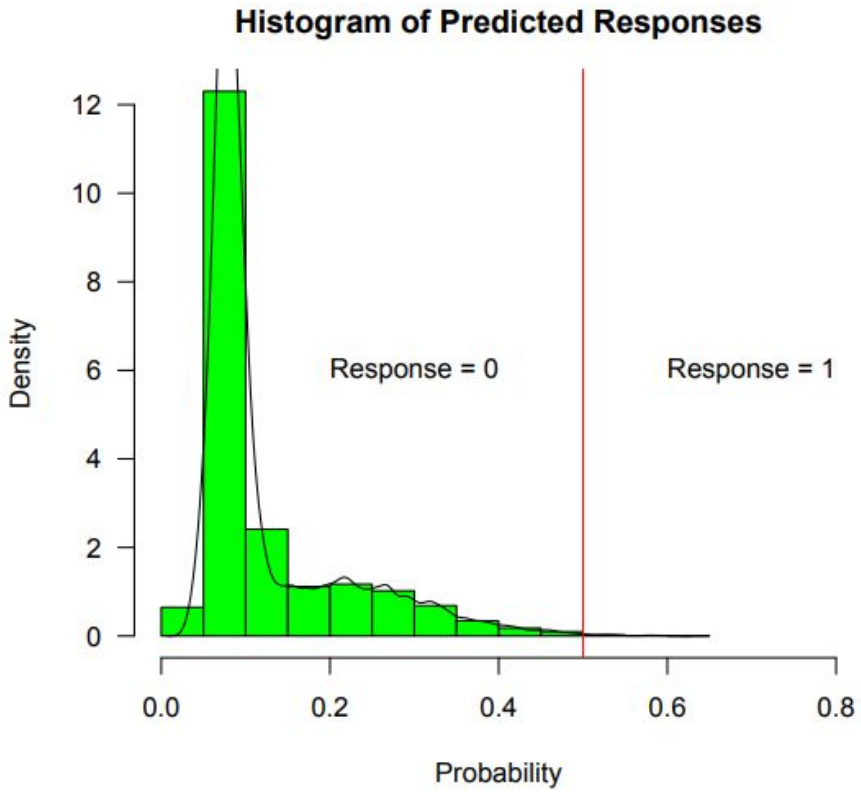


Results

Log Loss = 0.5640

Majority of Predicted Responses are 0

Computationally Intensive



Conclusion

Further Research into Methods for Imbalanced Datasets.

Increased Computational Power ≠ Improved Performance

Dimensional Reduction Techniques