**Introduction**

* **Initial Dataset Overview**: This report covers a dataset used for AI Neural Net training for attempting to predict the likelihood of positive reactions to bank telemarketing [1]. The report also attempted to compare to a linear regression but provided no methodology and achieved a terrible performance which I found unlikely so my work was aimed at producing a more realistic classical statistics metric. The data set consisted of 41188 entries, starting with 21 variables which were a mix of categorical, numerical and Booleans with the output being a Boolean with 4640 positive responses.
* **Body of Work**: After running an initial binomial GLM I found that this performed very well at predicting positive outcomes so I began reducing the dataset to determine the amount of data required. These reductions focus on maintaining model performance relative to the AI's top predictions, which had a result that when selecting the top 50% of most likely to respond yes the Neural Net caught 78% of the actual positive responses. At each stage, variables were eliminated based on statistical significance, Z-values, and confidence intervals, progressively reducing complexity while aiming to retain predictive power.

**Variable Reduction**

**Full Model:**

* + The full model found 4611 of the 4640 total positives which is over 99% accuracy and needed only the top 18% of the contact list to match the 78% coverage of the Neural Network

**First Reduction**:

* + Variables such as euribor3m, pdays, campaign, emp.var.rate, and cons.price.idx are identified for removal based on low Z-values and insignificant confidence intervals (i.e., intervals not containing 0).
  + The model is trimmed to focus on a subset of variables with a reduction to 4613 4578
  + entries at 50% prediction accuracy and 3657 still at 18%, maintaining parity with the AI model.

**Second Reduction**:

* + Notably, default, housing, and loan variables are found irrelevant and removed.
  + After further variable pruning, the dataset is reduced to 4578 entries at 50% accuracy and 3631 at 18%, showing that further simplification is possible without loss in model fit as with half of the original predictors the top 50% is still over 99% coverage and needs only 18% of the calls to get 15% coverage.

**Third Reduction**:

* + Here, removing days and months results in failing to meet the 78% accuracy threshold and requiring a switch to the top 19% of predictions.

**Fourth Reduction**:

* + At this point, the model continues performing similarly to the deep-learning neural network using only 5 parameters education, default, duration, previous and consumer confidence index.
  + The final dataset has 4204 entries at 50%, requiring an increase to the top 29% of predictions to keep up with the Ais 78% coverage rate.
  + Any further reductions significantly impact prediction accuracy, suggesting this is the optimal balance between simplicity and predictive power.

**Model Coefficients**

Across reductions, the confidence intervals for predictors are found to show contributions to the model. Variables with statistically relevant and insignificant confidence intervals are eliminated in the first 3 reduction phase. Then for round 4 low Z value parameters were removed.

* **First Reduction:** Age, job, marital and poutcometrue
* **Second Reduction:** Housing and loan
* **Third Reduction:** Day, month and contact
* **Fourth Reduction:** Campaign, pdays, emp.var.rate, euribor3m and cons.price.idx
* **Final model:** Education, default, duration, previous and cons.conf.idx

**Overfitting**

The previous model trains and tests on the same dataset which upon testing caused overfitting. To handle this the data set was split into 75% used as the dataset for the GLM function and a 25% set the model was predicted onto. This causes a few errors as the models may not have all the same factor levels present as some have very few of a particular level e.g. the default parameters in the full over 40,000 entries only have 3 yes values so when splitting test and train data none where present in train so the predictive GLM model couldn’t handle these new parameters. This is a significant advantage in training a Neural Net for this application but with a sufficiently prepared and populated dataset, the GLM remains a better predictor as even while removing the insufficiently populated parameters which include cons.conf.idx one of the identified most important parameters to the model as well as poutcome, cons.price.idx and nr.employedand it still achieved 85% coverage in the top 50% predicted.

**Conclusion**

The Binomial GLM with 5 parameters significantly outperforms a deep-learning model with 22 parameters in terms of simplicity and interpretability, without sacrificing much predictive power. This analysis highlights the value of variable selection and reduction in maintaining model accuracy while avoiding overfitting or model complexity. Additionally even when reducing the number of data points and variables and accounting for overfitting the GLM still achieves 7% more positive response coverage in the top 50% than the Neural Network.

**References**

[1] S. Moro, P. Rita, and P. Cortez. "Bank Marketing," UCI Machine Learning Repository, 2014. [Online]. Available: https://doi.org/10.24432/C5K306.