### **Predictive Modeling Midterm Report**

#### **Part 1: Generalized Method of Moments (GMM) with Delta Bias Term**

**Objective** The purpose here was to evaluate the presence of a bias term-delta- in a GMM model for Stock Change. This model used instrumental variables (IVs) to address possible endogeneity between Inventory Turnover and error terms, with data based on the stock return information of small retailers.An industry expert has suggested that bias exists within the moment conditions of these instrumental variables, necessitating the testing of a delta term in the GMM model. "Current Ratio," "Quick Ratio," and "Debt Asset Ratio" served as instrumental variables to address endogeneity concerns. Including delta in the moment conditions allowed for testing the presence of any unobserved variable bias, providing a check on the industry expert’s claim regarding potential bias.

**Results/Interpretation** The estimated delta coefficient of -0.0003478, with a p-value of 0.8951, did not show statistically significant evidence of bias (where p-value > 0.05). This suggests that the chosen IVs address endogeneity quite effectively, which actively allows the model to perform well without needing more bias correction. Since there was no evidence supporting the expert’s claim, the GMM setup seems to be good enough for analyzing stock returns in small retailers without any additional adjustments for bias. Guidelines for choosing IVs highlight their purpose in capturing key relationships without adding extra complexity, which could otherwise lead to more variation in the model.

#### **Part 2: Logistic Regression for Credit Rating Classification**

**Objective** Part 2 focused on developing a logistic regression model to classify "Credit Rating" as Positive or Negative, in a context where a global bank seeks to expand into small cities and villages. The bank’s challenge in these areas is that many potential customers do not have a bank, and instead may rely on alternative payment methods or informal loans. The bank has a goal to automate credit application assessments to attract new customers and reduce the traditionally long processing times that may steer away potential applicants.

**Data Preprocessing**

Categorical encoding was done, wherethe target variable, "Credit Rating," was encoded as 1 for Positive and 0 for Negative, aligning with binary classification practices. As well as this, feature engineering was of importance. "Marital Status" was one-hot encoded, while ordinal features, such as "Monthly Income" and "Requested Credit Amount," were ordinally encoded. This approach makes sure each type of variable is encoded correctly, keeping the order for ordinal data and treating nominal data as separate categories.

**Threshold Analysis** At a standard threshold of 0.5, the model achieved a high recall of 1.00 for approvals (class 1), but showed zero precision and F1 scores for non-approvals (class 0), which may indicate a bias toward predicting approvals. In classification contexts, it is essential to balance precision and recall, as unbalanced predictions can result in skewed decision outcomes. The observed imbalance here highlights the model’s initial favoring of approvals, which has the potential to misalign with the bank’s goal of careful customer selection in underserved regions.To meet the bank’s requirement of limiting approvals to 15%, the threshold was adjusted to the 85th percentile of approval probabilities. This adjustment resulted in a recall of 0.85 for non-approvals (class 0) and 0.15 for approvals (class 1), which meant non-approval predictions were prioritized to minimize credit risk while lowering the overall approval rate. However, the accuracy decreased to 0.25, underscoring a trade-off between sensitivity and accuracy.

**Theoretical Justification** The threshold adjustment aligns with principles of the bias-variance trade-off, where models are tuned to prioritize specific outcomes—in this case, limiting approvals in higher-risk small-city regions. This adjustment supports the bank’s expansion strategy by prioritizing recall for non-approvals, which means limiting unqualified approvals and reducing risk in these new markets.

#### **Conclusion and Core Business Implications**

1. **GMM Model Findings** The lack of statistical significance for delta indicates that the instrumental variables adequately address endogeneity, resulting in reliable coefficient estimates without additional complexity. This conclusion supports the robustness of the model for analyzing stock returns in small retailers, as the industry expert’s claim of bias was not statistically justified.
2. **Logistic Regression Model Findings and Practical Considerations** The adjusted logistic regression model meets the bank’s requirement by limiting approvals to 15%, which effectively addresses risk management goals by restricting approvals in small cities where traditional banking services are scarce. Although this conservative threshold may result in some qualified applicants being denied, it aligns with the bank’s business objective to minimize credit risk while expanding into new, less competitive regions. This trade-off highlights the importance of tuning models to meet specific business needs, especially in scenarios with high stakes.

Future improvements could include exploring alternative models, such as ensemble methods, to further balance sensitivity and precision, supporting decision making which is risk aware- this is important for the bank’s objectives for expansion into underserved regions.