Predicting a Customer’s Propensity to Join the Nile.com Loyalty Program

**Intro:**

The goal of this project was to develop a model that could predict if a customer would likely join Nile.com’s loyalty program by using historical data on customer demographics and previous orders. The target variable included in the dataset was a binary categorical field distinguishing members from non-members.

**Data Exploration:**

The first step was to explore the available data using descriptive statistics and visualization. Several important characteristics of the data were identified at this stage. The data had a strong class imbalance. Only 17% of the total customers were members of the loyalty program. Continuous values including purchase amount, age and days since last purchase were all normally distributed and their distributions differ significantly when split between the two target classes. There was no significant difference in number of the categorical features classes when split between target classes.

**Modeling:**

Given the absence of signal from the categorical variables, the initial feature set only included age, purchase amount and days since last purchase. Data was normalized and split between train (75%) and test (25%) sets. Logistic regression, SVM and Gradient boost models were tested using the sklearn library for Python. Model performance was evaluated using f1 scores and ROC curve. The best output using this dataset was F1 scores of 0.92 (not members) and 0.50 (members) and an AUC of .68.

Based on these results, a new feature set was created using a random subset of the dominant class of equal size to the minority class. Re-running the same models using this feature set, the F1 scores were both approximately equal for the two classes (around 0.8) and the AUC value went up to 0.8. Results were comparable for all three models.

**Conclusions:**

Based on these results, the class imbalance had a significant impact on the predictive ability of all models tried. With the original feature set, overall accuracy for logistic regression was 86% which seems mostly due to the model overwhelmingly predicting the dominant class. By taking a subsample of the dominant class, the models were more effective at predicting what had been the minority class.