Predicting Offer Acceptance Using Random Forests

J. Fuest | Siemens Advanta Consulting Analytics Cup

SIEMENS

Situation and Approach

Task

Build a model that predicts whether a customer will accept a sales offer:

- Binary classification problem
- Evaluated on balanced accuracy of predictions on test data set

Data

1. Past offers (30 000 rows)
2. Information on customers (8000 rows)
3. Geographical data on sales (100 rows)

Challenge

1. Poor data quality and labelling
2. Highly skewed training data with 80% accepted offers, which can lead to lack of sensitivity to rejected offers in model

Data Wrangling

Enhanced dataset by computing new, useful variables from raw data, such as discounts and margin

Model Choice

Random Forest using R's ranger library due to following benefits:

- Strong empirical track record (Lessman and Voss, 2010)
- Particularly suitable for high-dimensional data
- Good out-of-the-box performance

Model Training

Tested three different methods for dealing with skewed data against using original data to avoid bias

Deep dive on the right

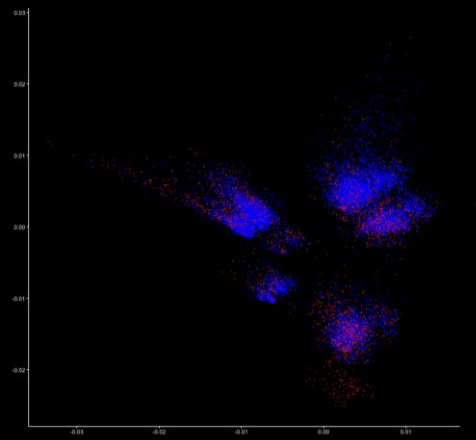
Tuning

Implemented grid searches to find the optimal hyperparameters for each method

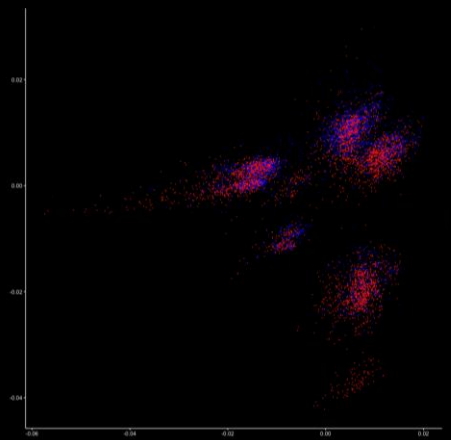
SIEMENS

PCA Visualization of Techniques Used to Balance Data

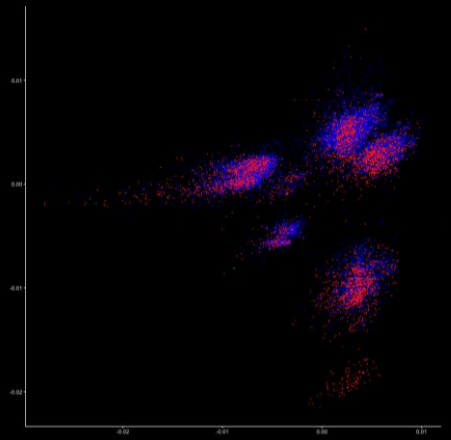
Original Data



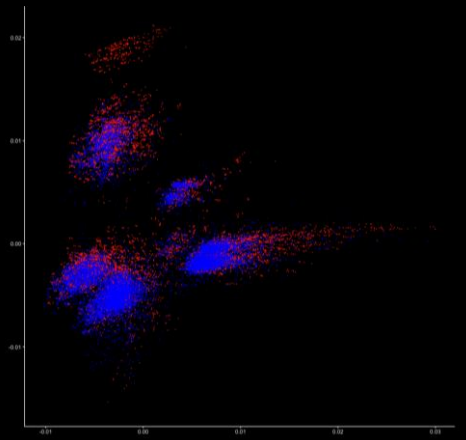
Random Undersampling



Random Oversampling



SMOTE



- Rejected Offers
- Accepted Offers

Results and Outlook

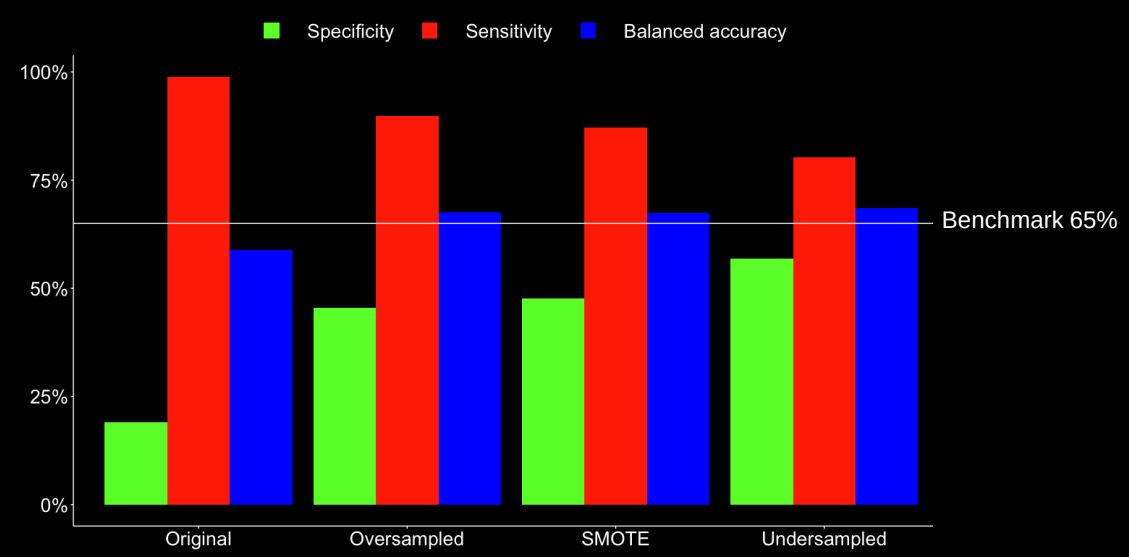
Results

1. All three methods met the desired balanced accuracy benchmark for achieving the top grade
2. Newly created variables were among the most important

Outlook

1. Significant gap remains between sensitivity and specificity, showing room for improvement
2. Combination of models could improve accuracy (ensemble learning)
3. Variable selection could further improve accuracy

1



2

SIEMENS