Santander Customer Transaction Prediction

Team 21
Derartu Dinku, Kelly Field, Sy Lax, Kishan Patel

Problem Statement

Challenge

The lack of insights into customer behavior affects Santander’s ability to provide proactive customer service and target marketing impacting acquisition costs and customer lifetime value.

Solution

Build a predictive model that identifies which customers will make a transaction in the future regardless of the amount of money transacted - and generate insights into the features most important in identifying customer transactions.

Data Description

- 200 anonymized numeric features
- ID column
- Binary ‘target’ response column
  - 0 = No Purchase; 1 = Purchase
- Train Data - 200,000x202
- Test Data - 200,000x201
  - No ‘target’ variable with actuals

Data Preview

Train Set

ID_code	target	var_0	var_1	var_2	var_3	var_4	....	var_198	var_199
train_0	0	8.9255	-6.7863	11.9081	5.093	11.4607		12.7803	-1.0914
train_1	0	11.5006	-4.1473	13.8588	5.389	12.3622		18.356	1.9518
train_2	0	8.6093	-2.7457	12.0805	7.8928	10.5825		14.7222	0.3965
train_3	0	11.0604	-2.1518	8.9522	7.1957	12.5846		17.9697	-8.9996

Test Set

ID_code	var_0	var_1	var_2	var_3	var_4	....	var_198	var_199
test_0	11.0656	7.7798	12.9536	9.4292	11.4327		15.4722	-8.7197
test_1	8.5304	1.2543	11.3047	5.1858	9.1974		19.1293	-20.976
test_2	5.4827	-10.3581	10.1407	7.0479	10.2628		19.8956	-23.1794
test_3	8.5374	-1.3222	12.022	6.5749	8.8458		13.0168	-4.2108

Notebook Critique - What Makes them Stand Out?

- Clear layout walking through each step to arrive at final model
- Explanations of each step that allow for reproducibility
- Select Visuals that aid in data exploration
- Eliminated Synthetic Samples to Improve Score

Key Issues with Dataset

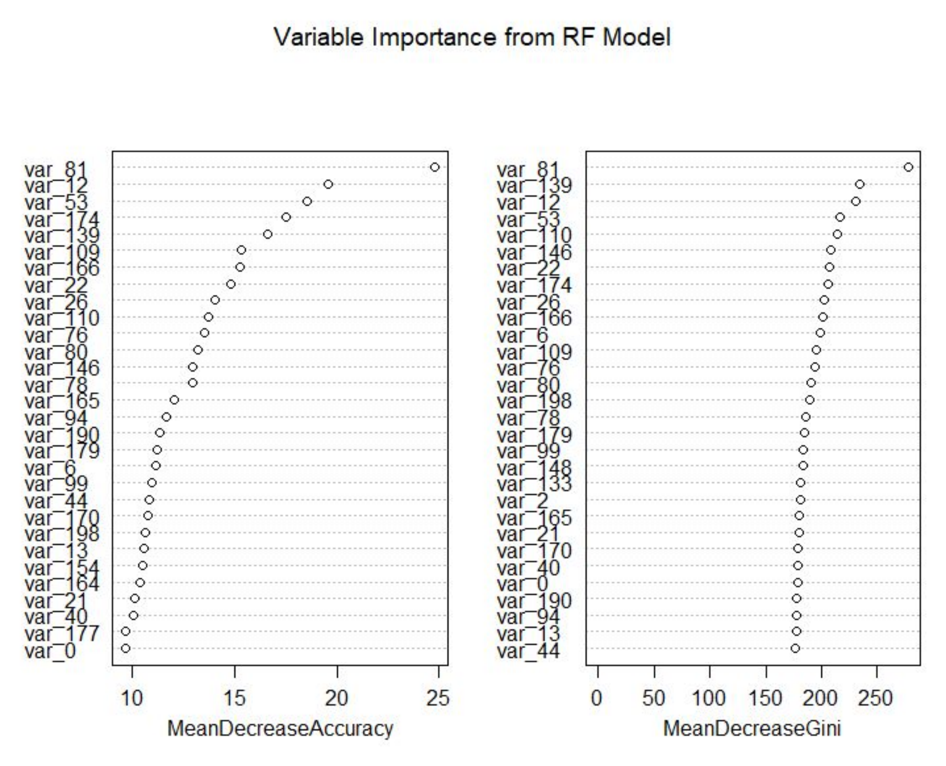
- Feature Selection
  - Data set containing large number of features
  - Deciding how to use each feature - lasso, ridge
- Imbalanced Target Column in Train Set
  - More 0 instances than 1's
  - Creating a model that handles the imbalance to combat bias towards one response

Exploratory Data Analysis

Target Class Proportions in Training Data

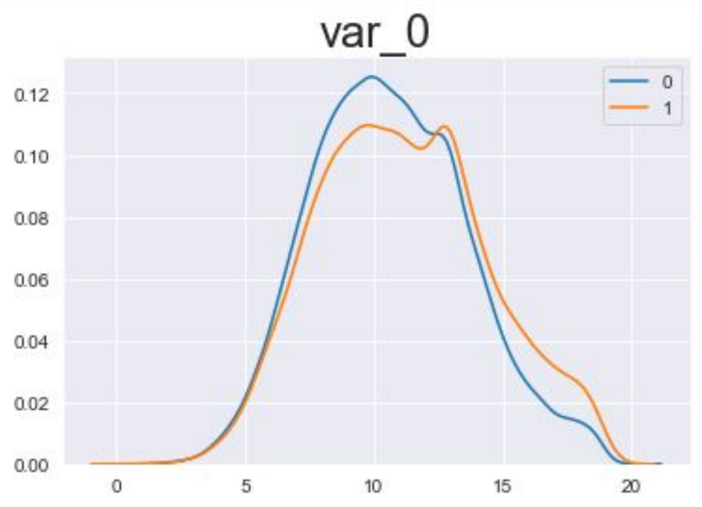
0 - 89.951%

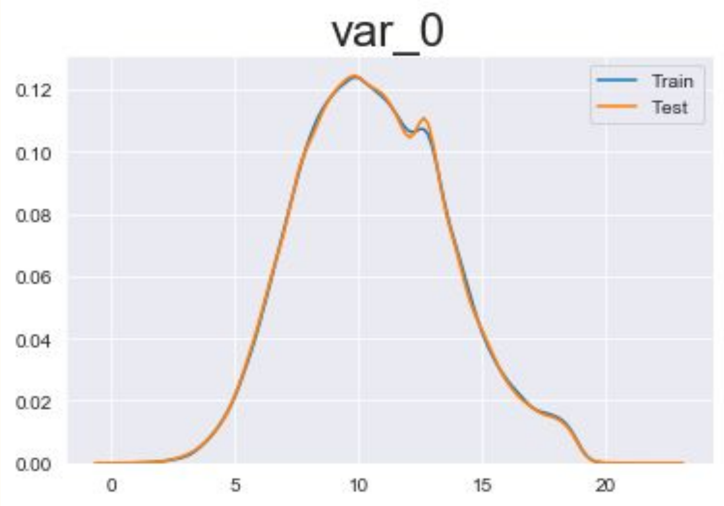
1 - 10.049%



Exploratory Data Analysis

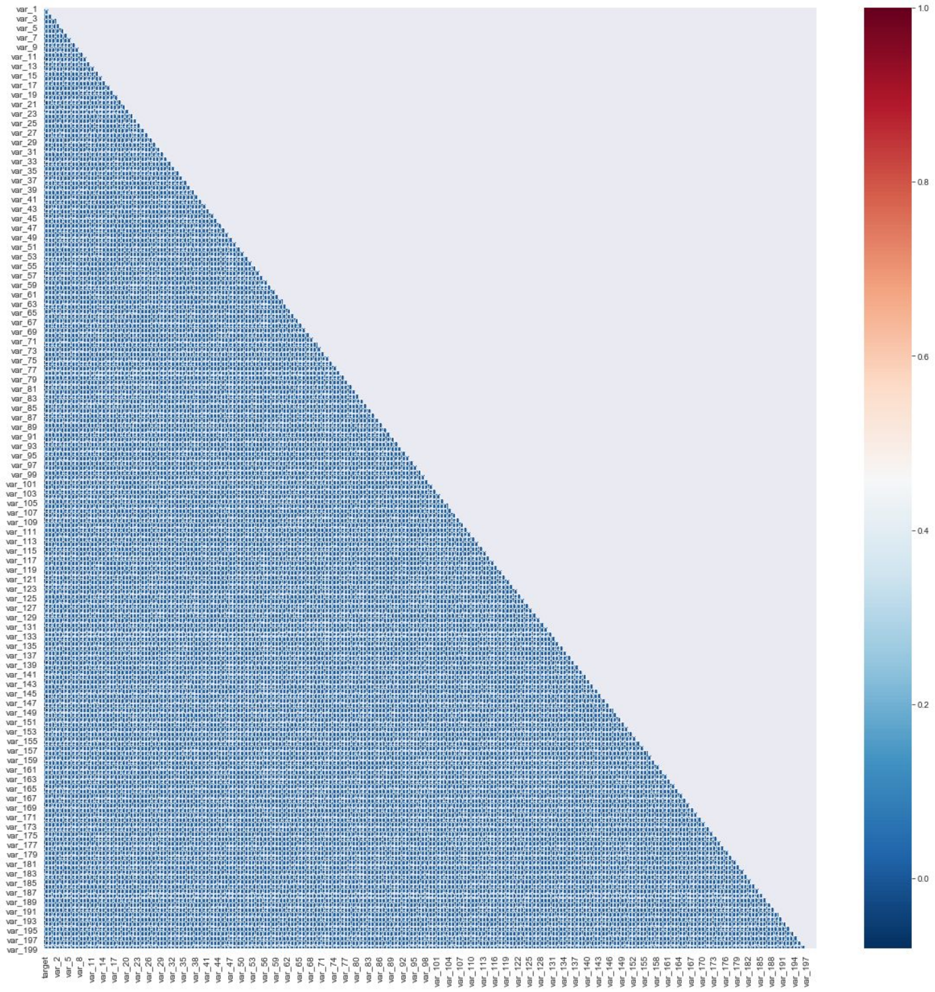
- All features are normally distributed.
  - Lead us to believe that the variables are the result of a PCA transformation





Correlation Heat Map

No correlation between
variables - independent



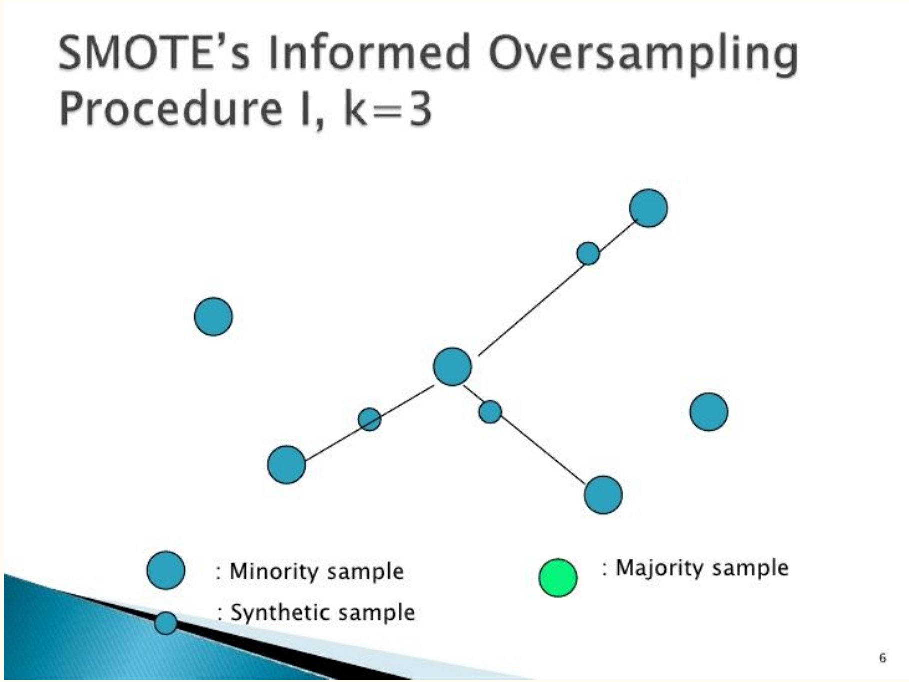
Training and Validation Process

- 80/20 split of the train data set
  - 160,000 training rows and 40,000 validation rows
  - Consistent samples in all models
- Tried sampling to deal with class imbalance
  - SMOTE and Under-sampling
- Fit and evaluated multiple models
  - Random Forest, Logistic Regression, SVM, Boosting, Lasso/Ridge and Naive Bayes

Balancing the Data

- Why is it important?
  - Predictions
  - Evaluation Metrics

- Methods
  - Down Sampling
  - SMOTE



Experimentation Results

No Sampling

Model	ROC AUC	Accuracy
Logistic Regression	0.629	0.915
Random Forest	0.5	0.9
Lasso	0.606	0.914
Ridge	0.626	0.915
Naive Bayes	0.8916	0.92215

SMOTE

Model	ROC AUC	Accuracy
Logistic Regression	0.784	0.783
Random Forest	0.599	0.633
Lasso	0.736	0.590
Ridge	0.738	0.596
Naive Bayes	0.524	0.0899

Under-Sampling

Model	ROC AUC	Accuracy
Logistic Regression	0.8633	0.780
Random Forest	0.870	0.776
GBM	0.872	0.782
Ridge	0.811	0.900
Naive Bayes	0.891	0.809

Final Model: Naive Bayes Classifier

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Model Assumptions

- Predictors are independent
- All predictors have an equal effect on the outcome

Model Results

- Accuracy - 0.92215
- ROC AUC - 0.8916
- Average 10-Fold CV ROC AUC Score - 0.889

ACTUALS

PREDICTIONS	0	1
0	35413	2521
1	593	1473

Reproducibility

Datasets

- https://www.kaggle.com/c/santander-customer-transaction-prediction/data

- Minor preprocessing steps to prepare the data

Reproducibility

Model: Naive Bayes Classifier

- Probabilistic Machine Learning model
- Based on Bayes Theorem
- Simple, effective and commonly used
- Does not require as much training ,non-sensitive to irrelevant features

Reproducibility

Code: FinalModel-Santander.R

- **Preprocessing Data**
  - performed using the train.csv provided since the test.csv lack the "target" column
- **Fit the model**
  - naiveBayes() function is used to fit the model
  - Make class predictions using the naïve bayes fit model using the test set
- **Evaluating the model performance**
  - Construct a confusion Matrix
  - Calculate overall accuracy rate, ROC Area Under Curve, PRAUC
  - 10-fold cross validation