Predicting a Customer's Future Transactions

Sheena Cook

Table of Contents

Applications of Data Science in Finance

Decision Making

Customer Analytics

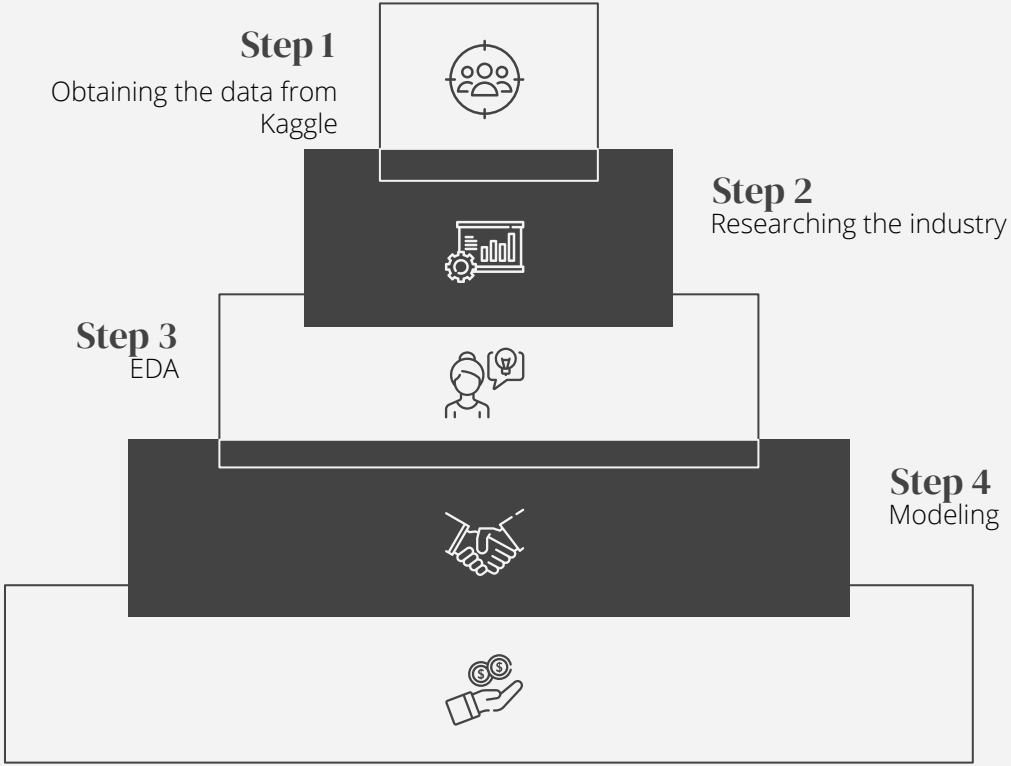
Fraud Detection

Customer Data Management

Problem Statement

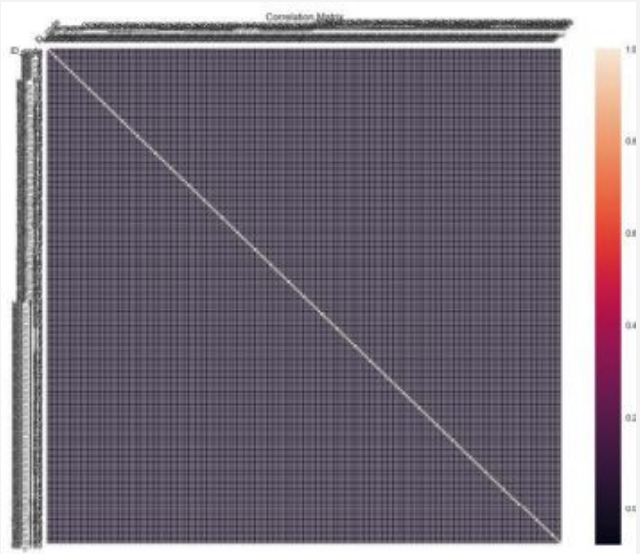
Can a model be created for Santander bank that can identify a customer who will make a specific transaction in the future.

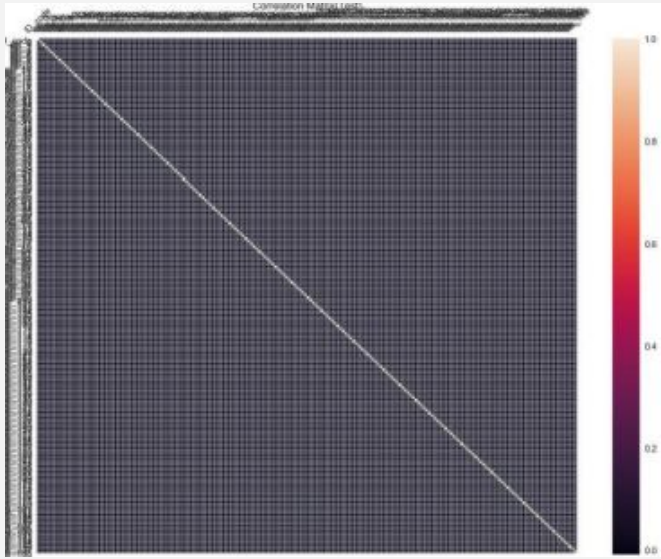
Workflow


Step 1
Obtaining the data from Kaggle
Step 2
Researching the industry
Step 3
EDA
Step 4
Modeling
Step 5
Interpretations & Next Steps

EDA

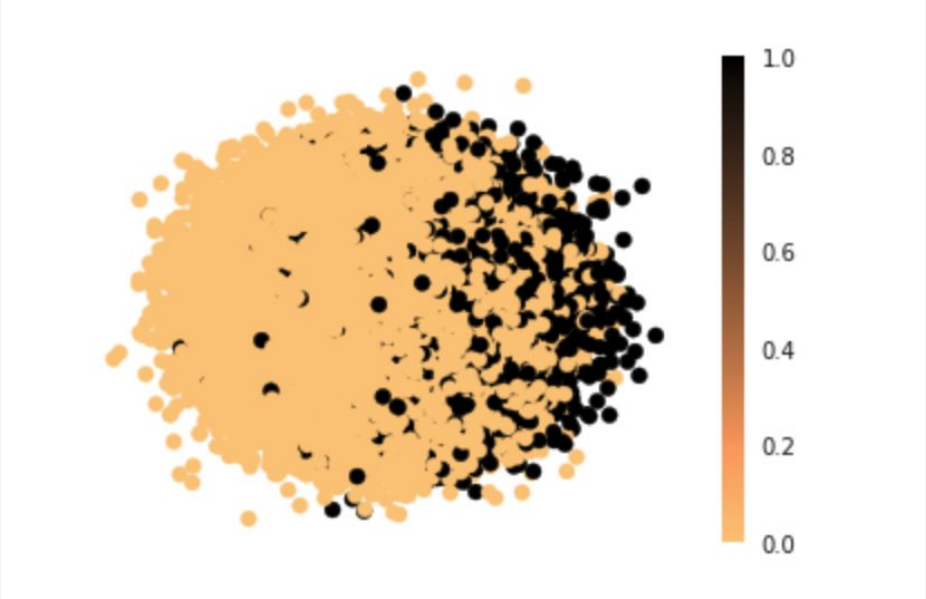
Feature Correlation





Data reflects no feature correlation between train or test

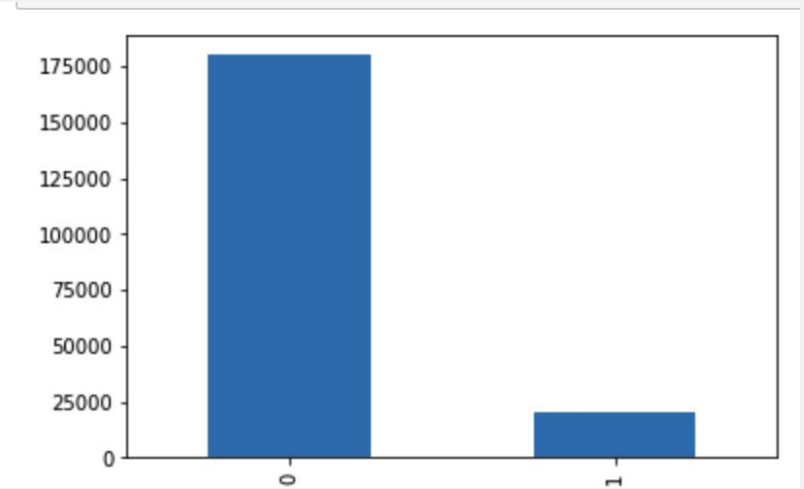
PCA



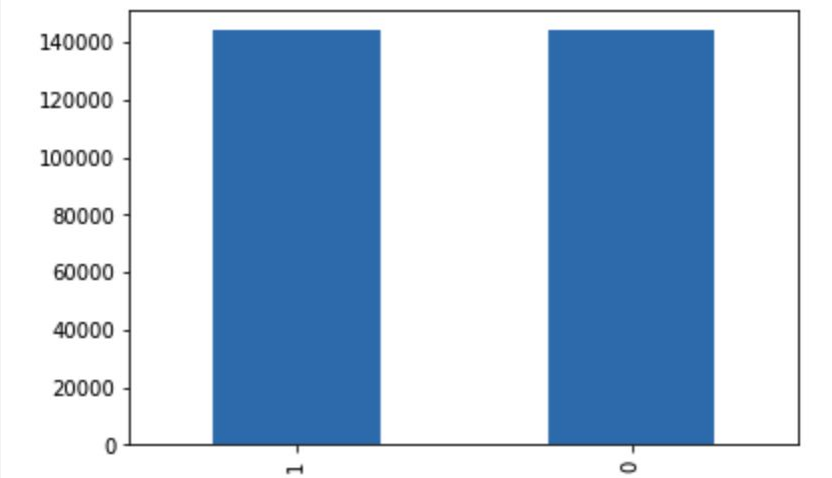
PCA aims for dimensionality reduction

Class Imbalance

Target Variable



Target Variable SMOTE



Majority class .89, minority .10 before application of SMOTE.

Modeling

Supervised Models w/ SMOTE

	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	Train: .79 Test: .79	.29	.76	.42	.86
Random Forest	.Train: 1 Test: .89	.25	.03	.05	.70
Naive Bayes	.Train: .86 Test: .87	.25	.03	.05	.60

***Models above are attempting to score better than the null model of .50**

***Key Performance Metric is AUC as it measures the ability of the classifier to distinguish between classes.**

Unsupervised Models w/ SMOTE

	Accuracy	Precision	Recall	F1-Score	AUC
Neural Net - dropout	Train: .88 Test: .85	.35	.49	.41	.95
Neural Net - early stop	.Train: .89 Test: 1	.28	.62	.38	.80

***Models above are attempting to score better than the null model of .50**

***Key Performance Metric is AUC as it measures the ability of the classifier to distinguish between classes.**

Interpretation of Models

Supervised Models

Logistic Regression – outperformed the attempted supervised models with a .86 AUC Score. **Naive Bayes** AUC score performed poorly compared to the other supervised models because NB assumes that points close to the centroid of class are likely to be members of that class, which leads it to mislabel positive training points with features. This explains why it underperformed compared to Logistic regression, as Logistic Regression is only concerned with correctly classifying points so the signal from outliers is more influential. Logistic Regression is also known to outperform NB on larger datasets.

Unsupervised Models

Neural Network– to combat overfitting I used regularized techniques to fit a NN. The Dropout Neural Network returned the highest AUC score. Tuning additional hyperparameters to increase AUC score.

Take Away & Next Steps...

Attempting additional models.

Additional tuning of hyperparameters to current models to optimize AUC score.

Undersampling data to decrease run time.

Collecting additional data for customer segmentation since we've determined customers who will make a future transaction.

Thanks

Does anyone have any questions?