Module 3

Santander Kaggle Dataset

Chris Gudde

Project Overview

The task of this project is to create a machine learning model which can produce a binary classification (0 or 1) based on 200 predictors.

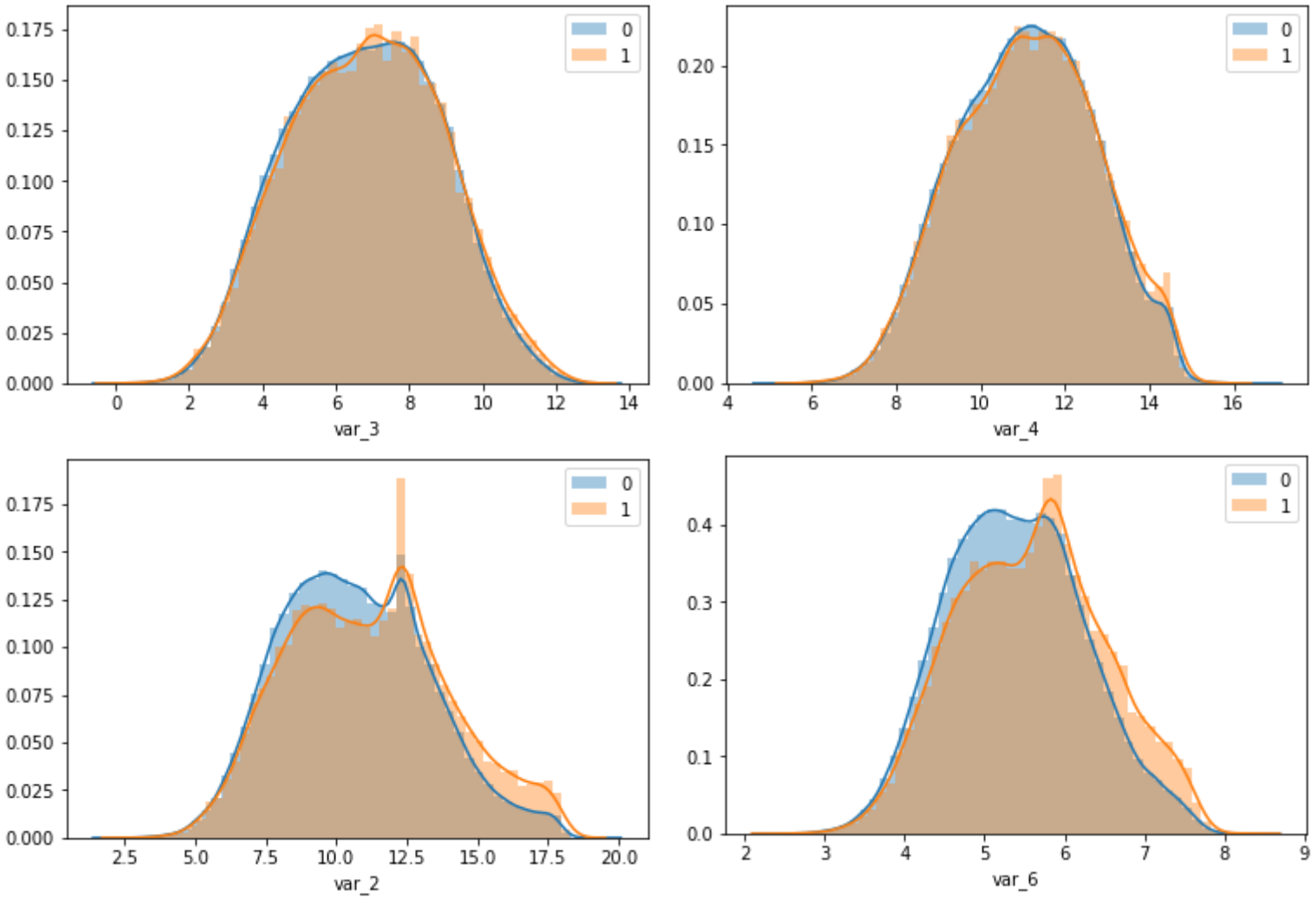
The dataset has 200,000 training records (rows) and the challenge is to predict 200,000 test records.

The test records do not have labels, to get the accuracy of the model with regards to the test records the predicted labels must be submitted to Kaggle

Project Framework

	Pass 1	Pass 2	Pass 3	Pass 4
Obtain	Import base data			
Scrub	Establish data quality			
Explore	Establish distributions	Normalise dataset		
Model	Create baseline model	Optimise Model	Optimise Model	Optimise Model
iNterpret		Submit to Kaggle	Submit to Kaggle	Submit to Kaggle

Exploratory Data Analysis

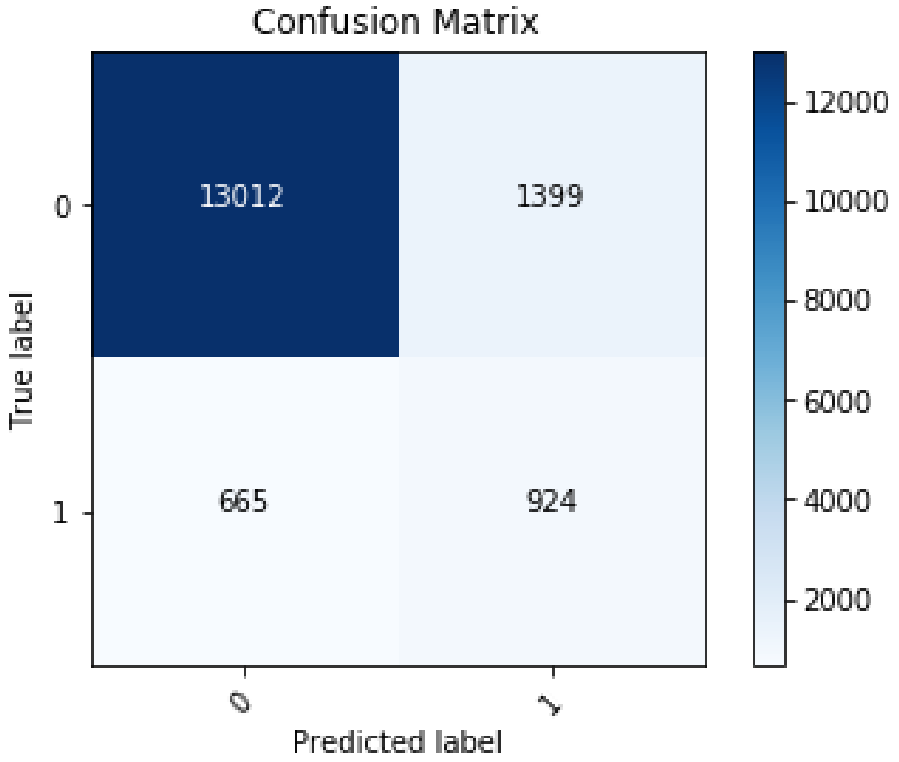


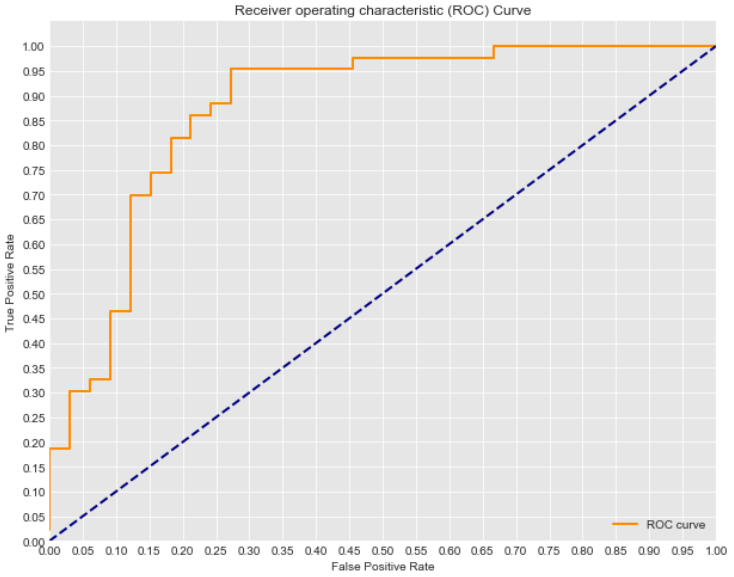
So why is this challenging?

- Large imbalance, only ~20,000 of the 400,000 records are positive
- Marginal variation in predictor variables
- Relatively large dataset (300MB)

Model Optimisation

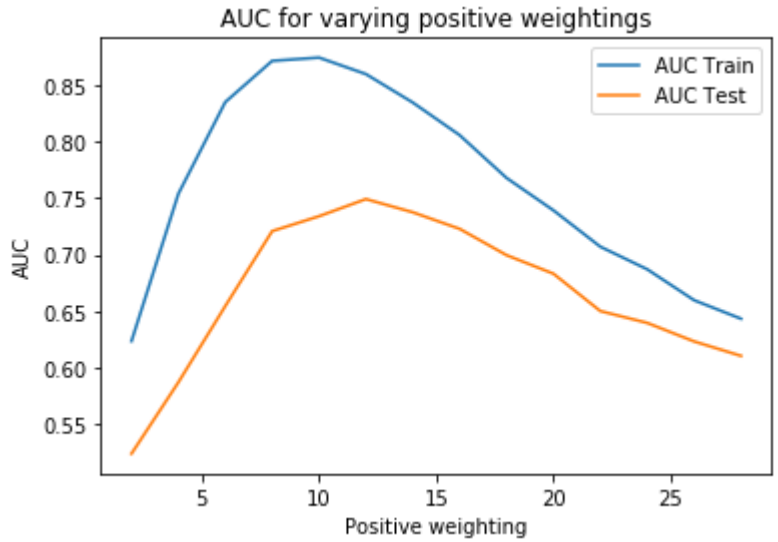
- The aim of the challenge is to get a develop a highly optimised model.
- The assessment criteria is Area Under the Curve (AUC)





Local Model Results

		Test	Train
1	Un-optimised Tree model	0.54	0.5
2	Tree model – imbalance roughly accounted for	0.74	0.58
3	Tree model – imbalance optimised	0.85	0.75



0.5 in a binary classification problem is equal to random chance

1.0 in a binary classification problem is 100% accuracy

Cloud Model Results

		Test	Train
1	Un-optimised Tree model	0.54	0.5
2	Tree model – imbalance roughly accounted for	0.74	0.58
3	Tree model – imbalance optimised	0.85	0.75
4	Tree model – multiple parameter optimisation	0.93	0.87
5	Tree model – multiple parameter optimisation (full)	0.91	0.90

Model: XG Boost

```
[0]     train-auc:0.552023      valid_data-auc:0.546484
Multiple eval metrics have been passed: 'valid_data-auc' will be used for early stopping.

Will train until valid_data-auc hasn't improved in 400 rounds.
[100]   train-auc:0.879516      valid_data-auc:0.829423
[200]   train-auc:0.91479       valid_data-auc:0.851685
[300]   train-auc:0.931898      valid_data-auc:0.860228
[400]   train-auc:0.943543      valid_data-auc:0.862927
[500]   train-auc:0.951019      valid_data-auc:0.864671
[600]   train-auc:0.95598       valid_data-auc:0.863158
[700]   train-auc:0.959963      valid_data-auc:0.865206
[800]   train-auc:0.963413      valid_data-auc:0.864257
[900]   train-auc:0.965218      valid_data-auc:0.864003
[1000]  train-auc:0.965218      valid_data-auc:0.864003
Stopping. Best iteration:
[695]   train-auc:0.959788      valid_data-auc:0.865413

Duration: 122.3475secs
```

0.5 in a binary classification problem is equal to random chance

1.0 in a binary classification problem is 100% accuracy

Further Work

- Further parameter optimisation
- Greater optimisation at full scale
- PCA / dimensionality reduction