Loop Q PRIZE Qualifying Challenge 2019

{ Michal Ostyk-Narbutt } Sapienza University of Rome

Introduction

The Willy Wonka chocolate factory produces the **Golden Scrumpilicious Candy Bar**, which packaged in a golden wrapper. However, sometimes during production the bars end up being green which is a defect, but also very alluring and well selling.
Given data from the production line from 60 Oompa-Loompa's, the goal is to predict the production of the **green bars**.
In this presentation I will go over the overview of the solution, its design criteria, and its strengths.

Methodology

The training pipeline(right), depicts the approach to the solution. At first, an SVM model was applied to the imbalanced data, but seeing that data is hard to marginalise in Figure 1, it gave really poor results. Hence, Random forests (RF) were used since model architecture reduces overfitting However, to improve results oversampling using SMOTE was peformed. RF precision and recall gave better results on a balanaced dataset. Note that in cross-validation (CV) with oversampling, only the training fold was oversampled.

Dataset

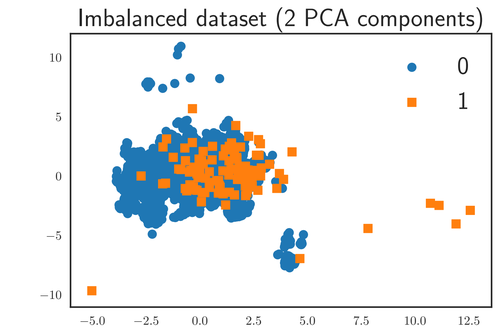


Figure 1: Normalised data with applied PCA dimensionality reduction. The plot shows how extreamly difficult it is to distinguish between golden and green bars.

The dataset given consits of 62 columns, of which 60 are scalar features given by each Oompa-Loompa, 1 is a time stamp, and the last is the target 'GREEN'. This task will consist of binary classification. However, the data is immensely imbalanced as green bar production is very rare.

Golden bars	Green bars
15704	102

Table 1: Gold vs Green bars in dataset

Feature engineering involved of decomposing the time stamp into 5 columns : "day", "dayWeek", "hour", "min", "month". Moreover, Features correlated with each other by a threshold $t = \pm 0.32$ were discarded. Hence, the training and testing was done on 22 features.

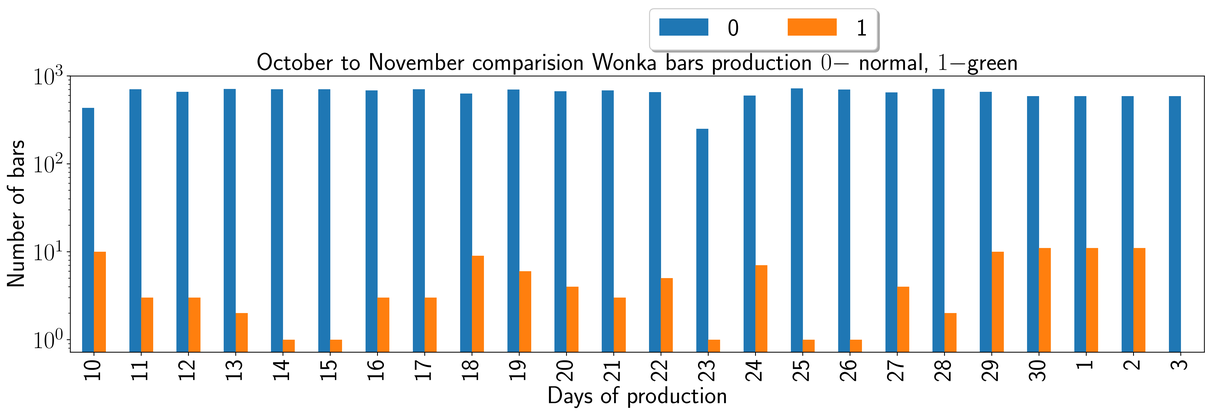
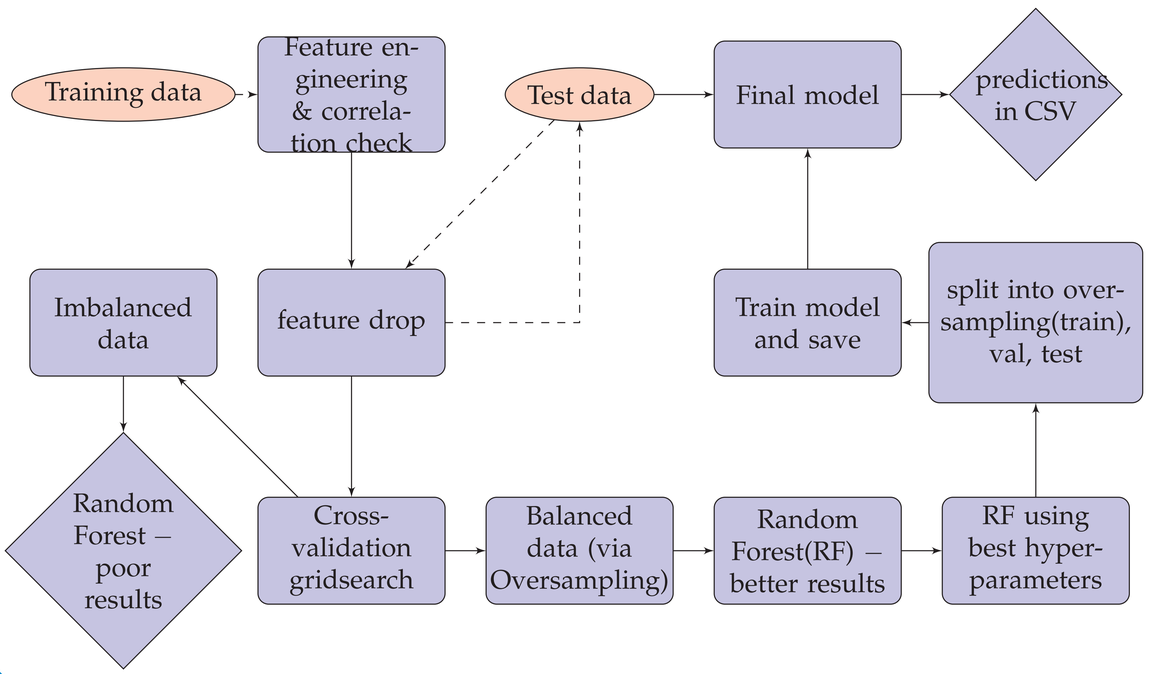


Figure 2: Analysis of production of gold(0) and green(1) chocolate bars through October to early November 2018

Training Pipeline



Results

The training consisted of a grid-search K-fold cross validation to find the best hyperparameters (Best-hp). Seeing that balanced data gave better results (Table 2), and using their Best-hp, training was performed on the entire dataset split into: train, validation, and test. Only train was oversampled.

Model Type	Accuracy	Precision	Recall	F1-score
CV Imblanced RF	0.9960±0.0006	0.91±0.05	0.41±0.08	0.9952±0.0008
CV Balanced RF	0.9947±0.0011	0.67±0.18	0.402±0.017	0.9941±0.0009
Best-hp *validation* macro average	0.99	1	0.75	0.83
Best-hp *test* macro average	0.99	0.96	0.68	0.76

Table 2: End results of training on balanced and imbalanced datasets + final training on the best hyperparamters.

Conclusions

Random forests are great estimators when their is a significant amount of collinerity on an imbalanced dataset. However, reducing that collinerity and oversampling the training set, results in better overall performance. In this task however, we are largly concered with recall than precision. This is because, it is far more costly to wrongly predict a green bar as gold, rather than vice-versa and hence letting a green bar simply slip by. The percentage of green bars in the training dataset is 0.65% meanwhile, the percentage on the given test dataset predicted by the best model is just over 1 percent. Hence to conclude, I reckon that the model should yield good results once tested by the judges, as greens are rare as potrayed in Figure 2.
Please note: logos in the top left and right belong to and were taken from <www.loopqprize.ai>