

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 performed. RF precision and recall gave better results on a balanced dataset. Note that in cross-validation (CV) with oversampling, only the training fold was oversampled.

DATASET

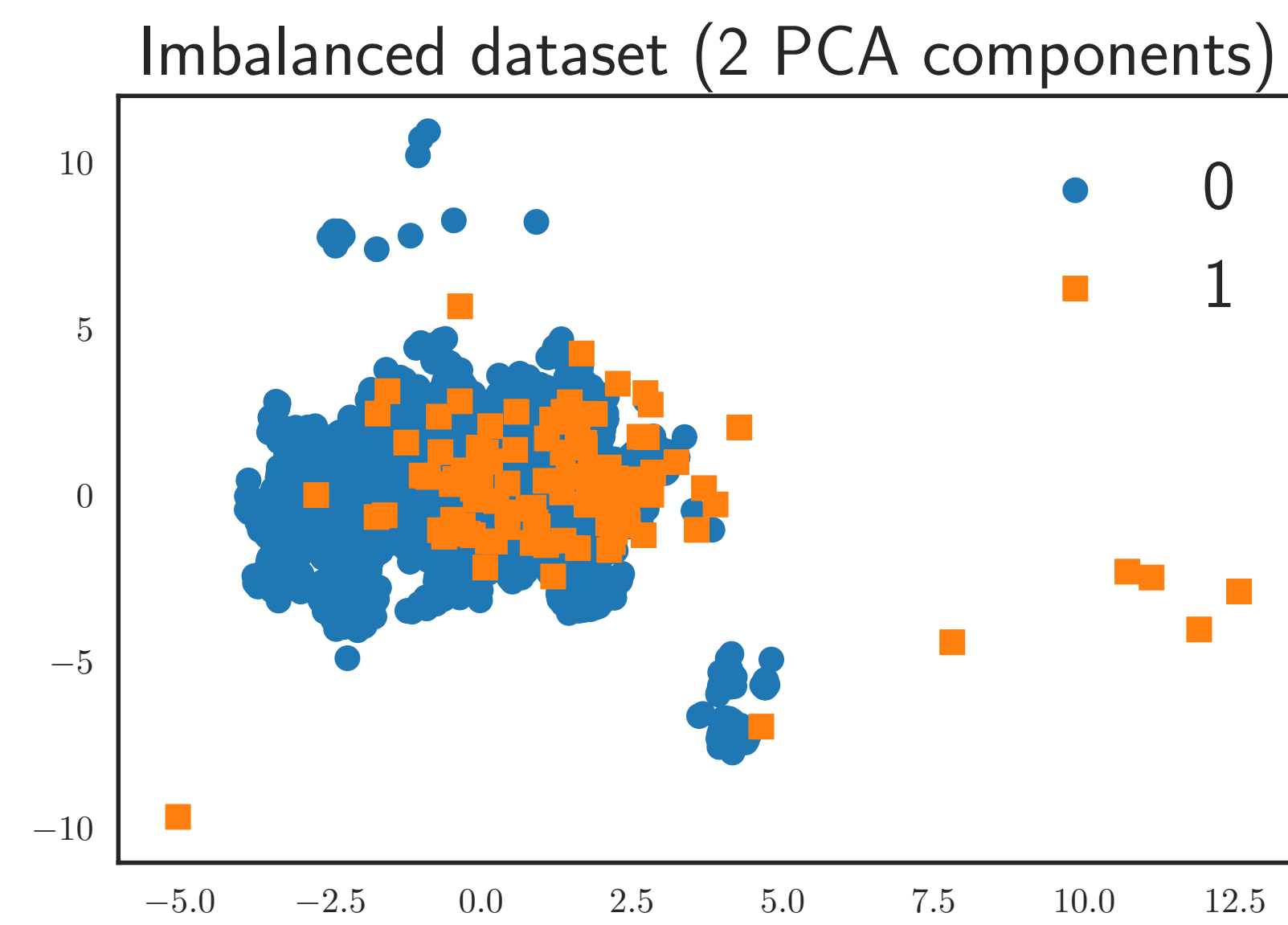


Figure 1: Normalised data with applied PCA dimensionality reduction. The plot shows how extremely difficult it is to distinguish between golden and green bars.

The dataset given consists 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.

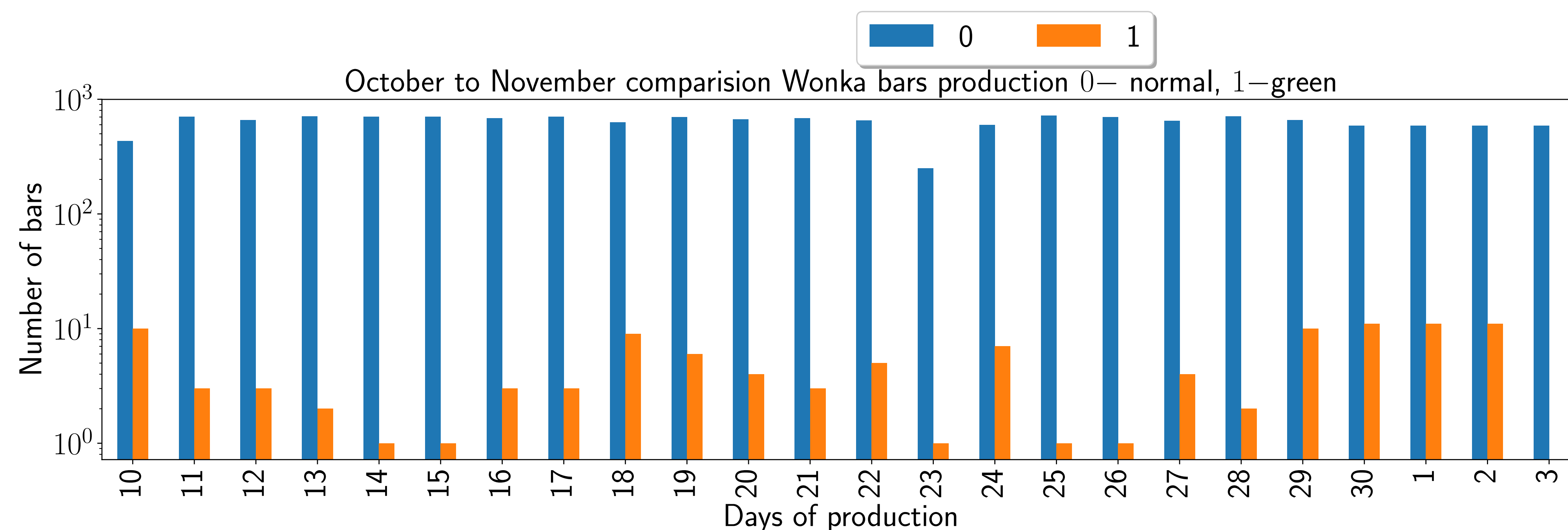
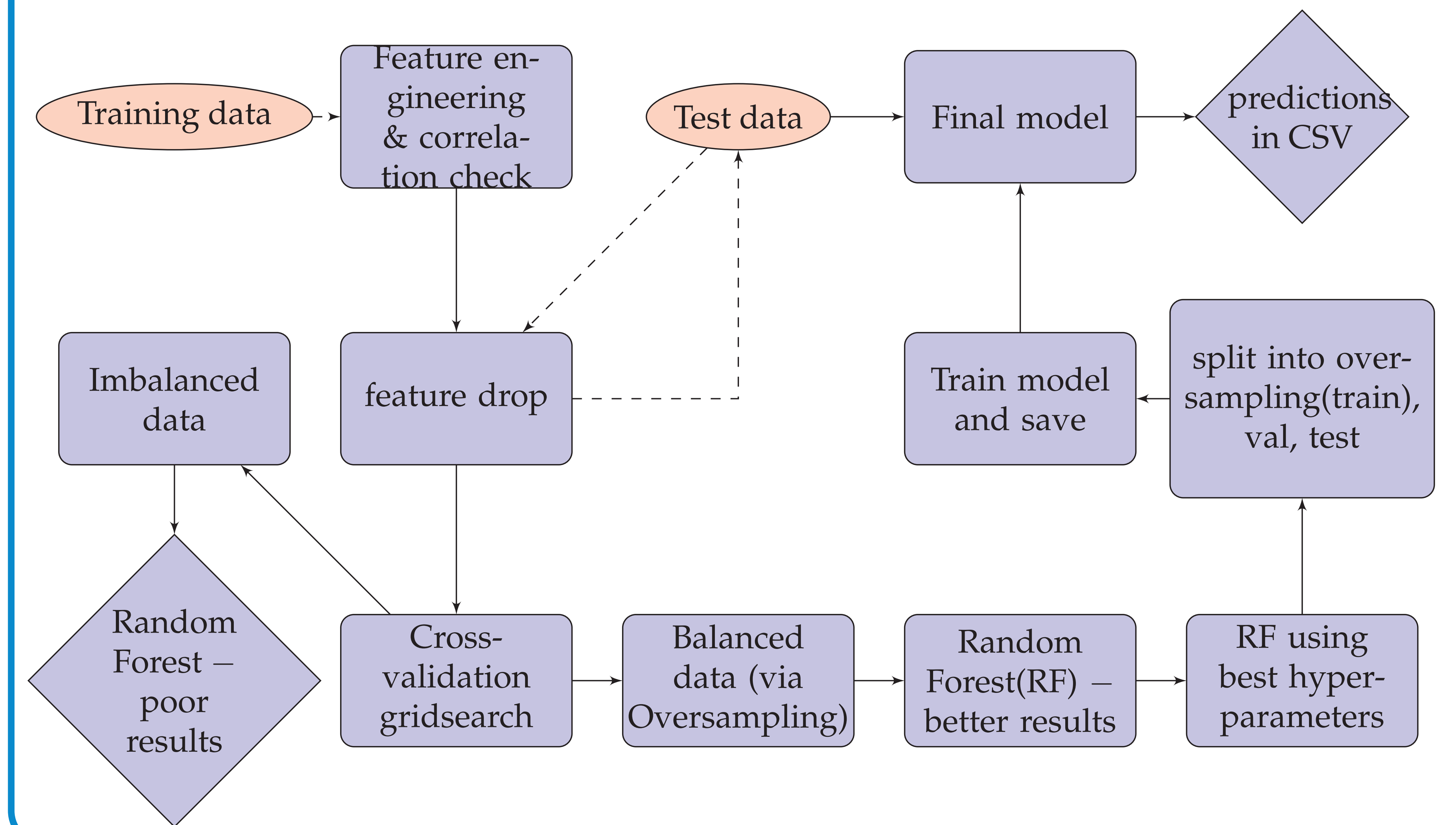


Figure 2: Analysis of production of golden (0) and green(1) chocolate bars through the month of October to early November 2018

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 Imbalanced 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 <i>validation</i> macro average	0.99	1	0.75	0.83
Best-hp <i>test</i> 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 hyperparameters.

CONCLUSIONS

Random forests are great estimators when there is a significant amount of collinearity on an imbalanced dataset. However, reducing that collinearity and oversampling the training set, results in better overall performance. In this task however, we are largely concerned with