Predicting New Particle Formation Events with machine Learning

Julia Sanders
Mikko Saukkoriipi
Bernardo Williams

University of Helsinki
Introduction to Machine Learning

December 11, 2020

Contents

Section 1

Overview of the Models

Introduction

Observation: Importance of stratified sampling

The data was split 60:20:20 into the training, validation and test sets.
Stratified Sampling over Class4

Train, validation test split			
	Train	Validation	Test
nonEvent	49.3%	50.0%	50.0%
Ia	7.1%	7.7%	7.7%
Ib	21.4%	21.2%	21.2%
II	22.1%	21.2%	21.2%

Table: Proportion of classes on each dataset

Binary Classifiers

Notes:

i. Validation might refer to validation or cross validation.

ii. Hyperparameter tuning was done by random grid search

Summary of binary models accuracies			
	Training	Validation	Test
Decision Tree	88%	84%	88%
Random Forest	100%	87%	88%
XGB	**100%**	**90%**	**87%**
KNN	85%	78%	80%
Logistic Regression	**87%**	**87%**	**89%**
Bayes (bestK)	81%	85%	87%
Bayes (PCA)	**84%**	**87%**	**93%**
SVM	98%	90%	83%

Table: Summary of tested binary models

Blended Model

The final blend of models chosen was **XGB, Naive Bayes and Logistic Regression**.

Accuracy	
Training	96.12 %
Validation	96.51 %
Test	91.86 %

Table: Blended Model Accuracy

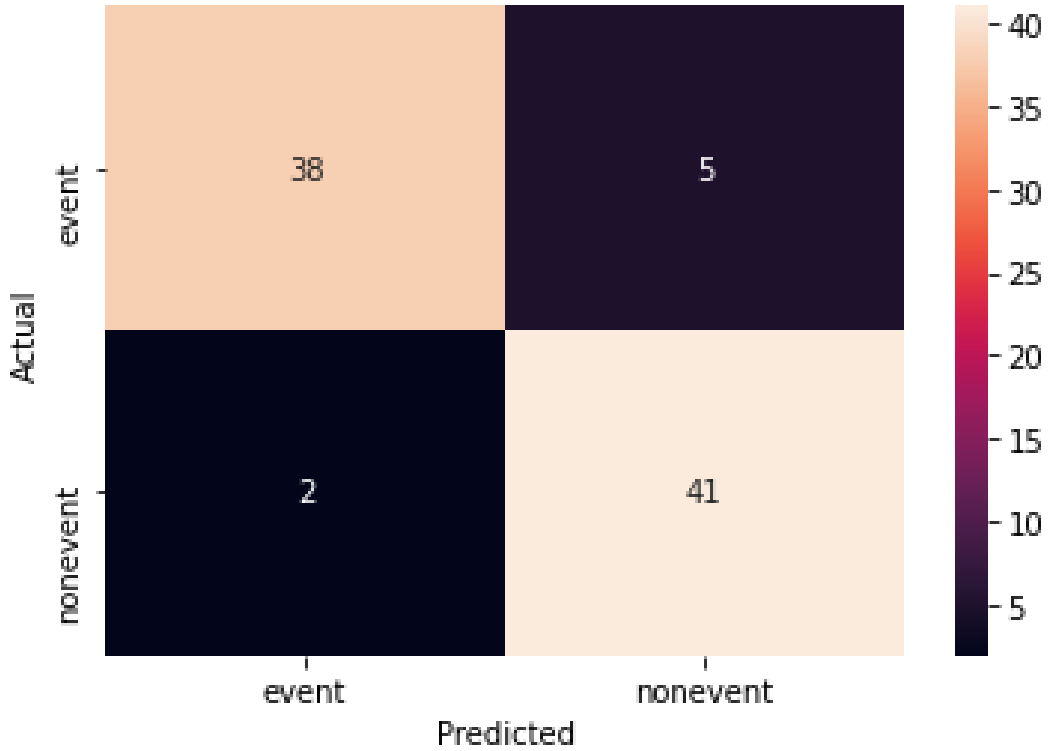

Fig.: Confusion Matrix Blend on Test

Multi-class Classifiers

Notes:

i. Validation might refer to validation or cross validation.

ii. Hyperparameter tuning was done by random grid search

Summary of multiclass models accuracies			
	Training	Validation	Test
Decision Tree	66%	64%	67%
Random Forest	100%	66%	72%
XGB	100%	70%	70%
KNN	66%	58%	58%
Bayes (bestK)	62%	64%	62%
Bayes (PCA)	69%	62%	65%
SVM	83%	69%	68%

Table: Summary of multiclass models accuracies

Multi-class Blended Model

The final blend of models chosen was **SVM, XGB and Naive Bayes**.

Accuracy	
Training	94.96 %
Validation	97.67 %
Test	69.77 %

Table: Multi-class Blended Model Accuracy

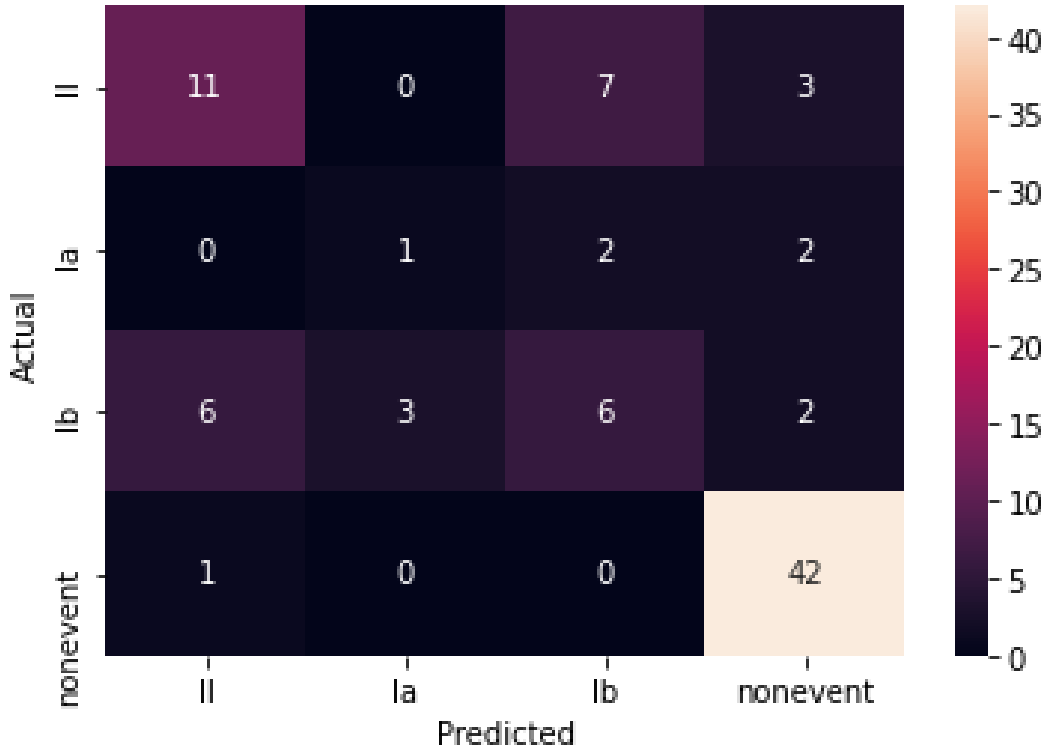

Fig.: Confusion Matrix Blend on Test

Section 2

Conclusion and Observations

Conclusions

- Why the model scored so highly on perplexity?
- Why was our estimate of binary accuracy so far off?

References

James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshira (2017)

An Introduction to Statistical Learning

Projects GitHub repository:

https://github.com/williwilliams3/TermProjectIML

Towards Data Science: Feature Selection Techniques in Machine Learning with Python.

https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

scikit-learn, Select Best K

https://scikit-learn.org/stable/modules/generated/sklearn.feature$_s$*election.SelectKBest.html*