Machine Learning:

Backorder prediction

University of Piraeus-
NCSR 'Demokritos'

Department of Digital Systems
MSc Artificial Intelligence

Evangelia Baou

Professor
Theodoros Giannakopoulos

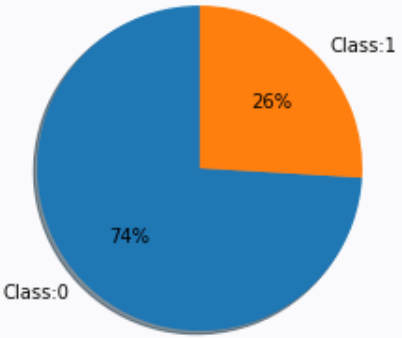
Athens, 2021

Problem Definition

- Prediction of orders which could not be fulfilled by the company(aka backorder).
- Why?
  - ✓ Increase customer's service level
  - ✓ Prevent over-stocking
- Binary classification problem

The dataset

- The dataset consists of 33356 samples and 22 attributes
- Types of data features: Numerical, categorical
- Negative values, null values
- Non-linear relationship between features and target variable



Data pre-processing

- Categorical to Boolean
- Delete nulls
- Replace negative values with zeros or nan (depending on the feature)

Classifiers' Selection

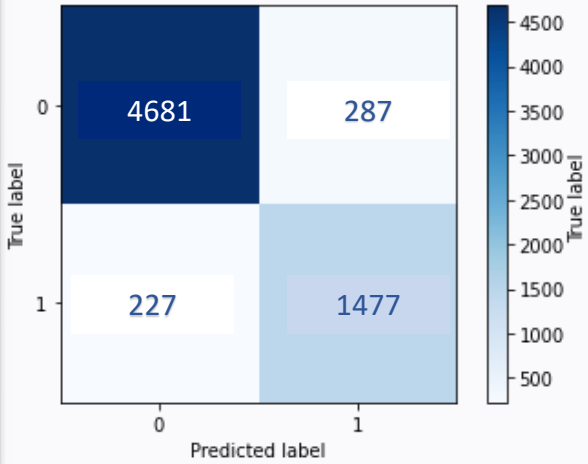
- Support Vector Machines(SVM)
- Random Forest
- KNN

Experiments' Steps

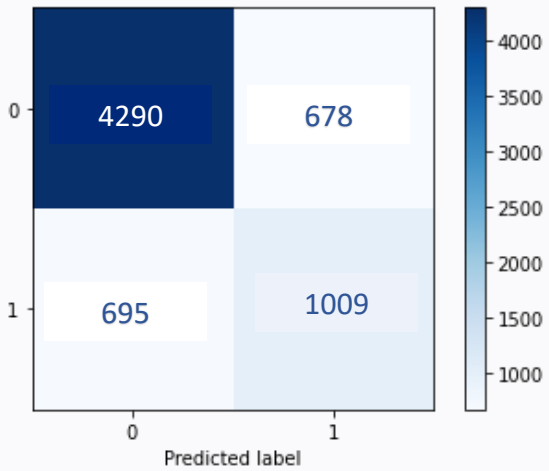
- Split dataset into train and test
- Impute and scale
- Fit models/predict classes
- Create confusion matrix and classification report for each one of them

Experiment Results

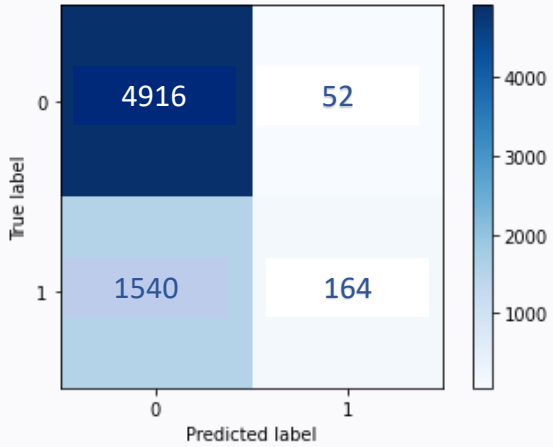
- Initially the algorithms are trained and tested with their default parameters and a SMOTE ratio=0.4



Random Forest Conf. Matrix



KNN Conf. Matrix



SVM Conf. Matrix

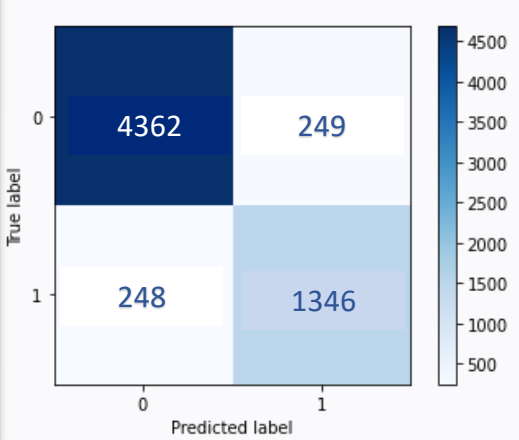
	F1-score	Precision	Recall	Bias	Variance
Random Forest	0.90	0.90	0.90	0.070	0.018
SVM	0.52	0.76	0.54	0.236	0.003
KNN	0.73	0.73	0.73	0.149	0.074

Outlier Removal and Classifiers' Hyparameter Tuning

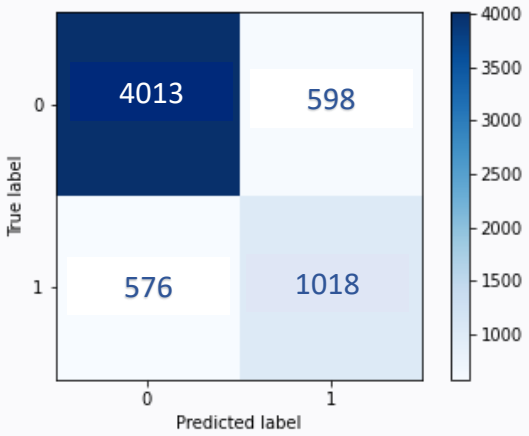
Outlier Removal

- How? Using z-score

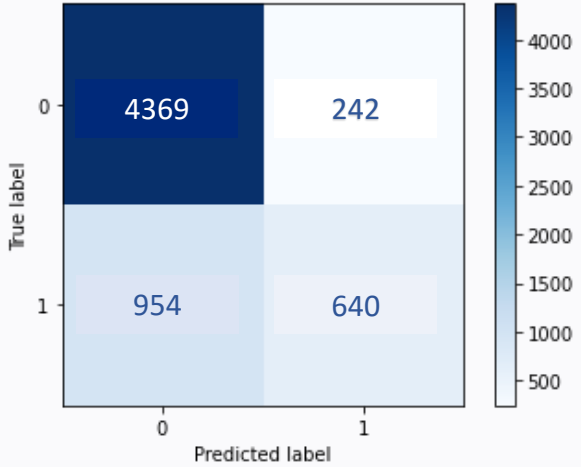
Hyparameter Tuning Results



Random Forest Conf. Matrix



KNN Conf. Matrix



SVM Conf. Matrix

	F1-score	Precision	Recall	Bias	Variance
Random Forest	0.90	0.90	0.90	0.070	0.019
SVM	0.70	0.77	0.67	0.180	0.016
KNN	0.75	0.75	0.75	0.136	0.069

Comparing Results

	F1-score	Precision	Recall	Bias	Variance
Random Forest	0%	0%	0%	0%	1%
SVM	7%	0%	5%	7%	34%
KNN	1%	1%	1%	2%	2%

- SVM indicates satisfying improvement
- KNN and Random Forest have no improvement

Conclusions

- Random forest performs better (in terms of f1-score) than SVM and KNN
- However, SVM indicates a really good performance for majority class (only 25 out of 4.968 were predicted as False positive) in contrast with Random Forest and KNN (287 and 678 out of 4.968 were predicted as False positive)

Future work

- Training each of the classifiers with different ratios of SMOTE
- Training SVM with feature selection
- What if we combine SVM and Random Forest?

THANK YOU