Advanced Topics in Machine Learning

Semester Project: Detection of APS Failure at Scania Trucks

Aleksei Baidarov

July 02, 2018

Table of Contents

Introduction

- Air Pressure System (APS) is critical for trucks (braking, gear changing)
- The cost of missing a faulty truck is 50 times higher than the cost of an unnecessary check of a truck
- Goal: minimization of APS maintenance costs:

$$Total\_Cost = 10 \cdot FP + 500 \cdot FN$$

Dataset Overview

1. Unbalanced (59 000 negative instances, 1 000 positive instances)

2. Many missing values (59 409 rows have at least one missing value, about 8% of values are missing in total)

3. High-dimensional (171 features)

Pre-Processing

Feature Selection:

- Univariate Feature Selection and Removing Features with Low Variance: no improvement $\rightarrow$ using all features

Handling missing values:

- Filling with mean value

Normalization:

- Using *StandardScaler* to make computations faster

Algorithms Used

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Support Vector Machines
5. LightGBM

General Approach

- Tools used: Python 3, Pandas, Scikit-learn
- *GridSearchCV* to find best parameters (Number of CVs = 5)
- *Recall* as Scoring function:

$$Recall = \frac{TP}{TP + FN}$$

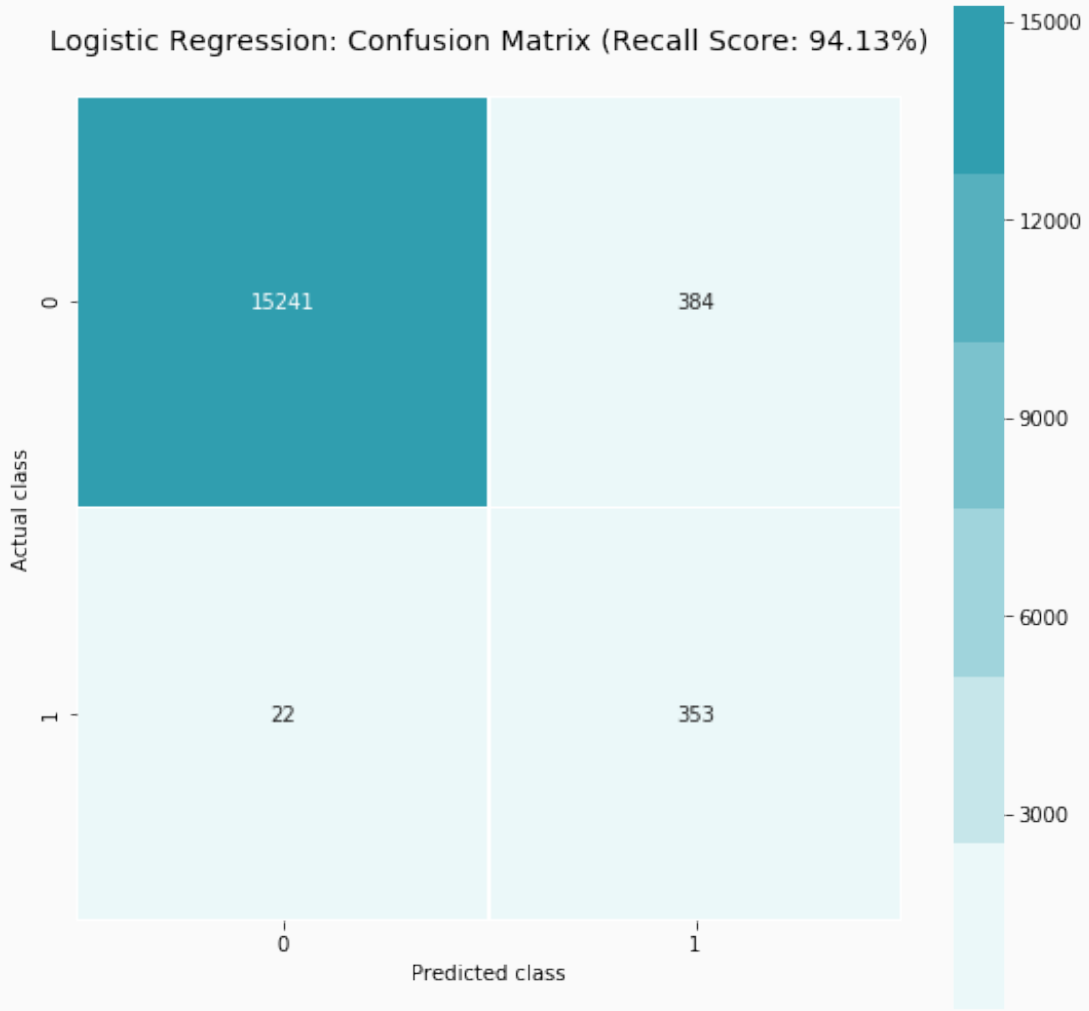
Results: Logistic Regression

Parameters:

- *C* : 0.0001, **0.001**, 0.01, 0.1, 1, 10, 100, 1000, 10000

Results:

- *Recall* = 94.13%
- *Total Costs* = 14 840



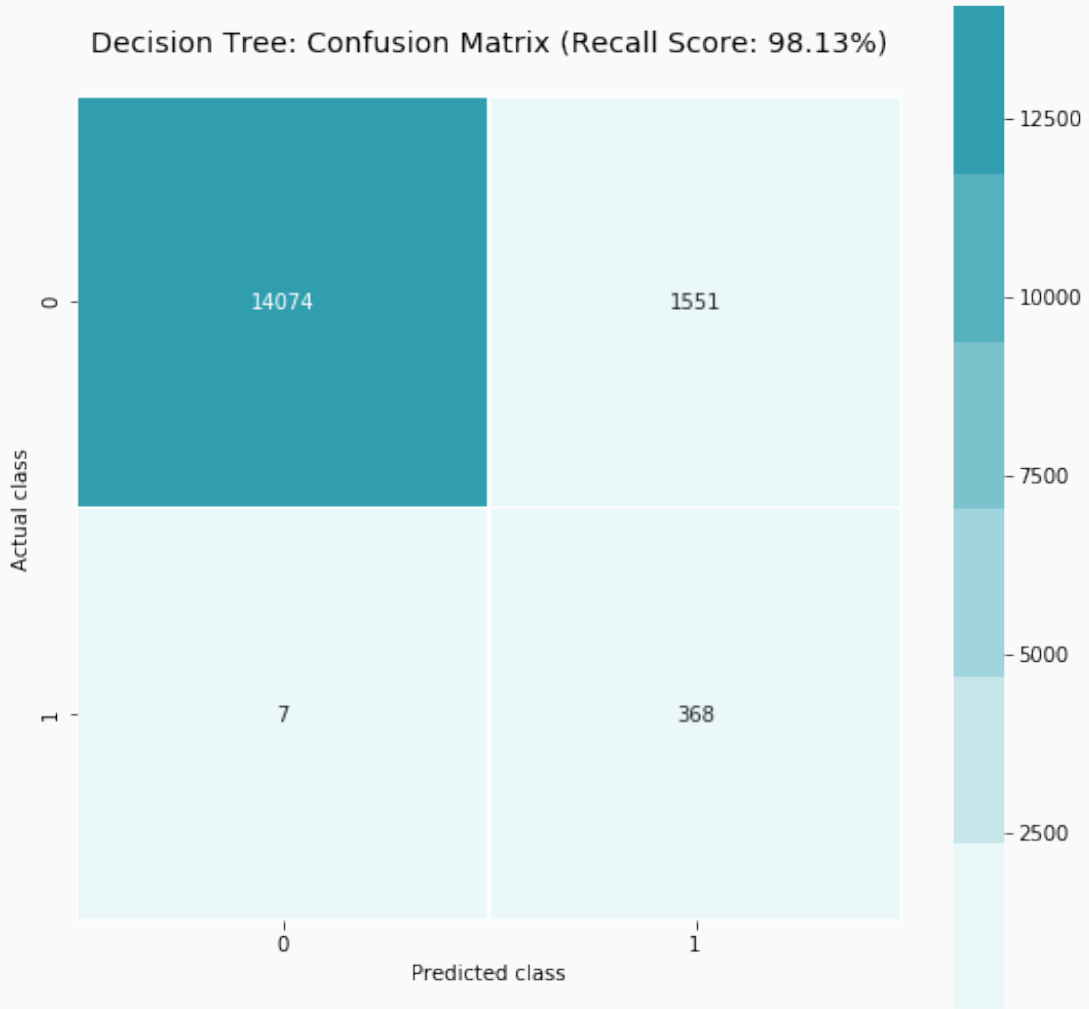
Results: Decision Tree

Parameters:

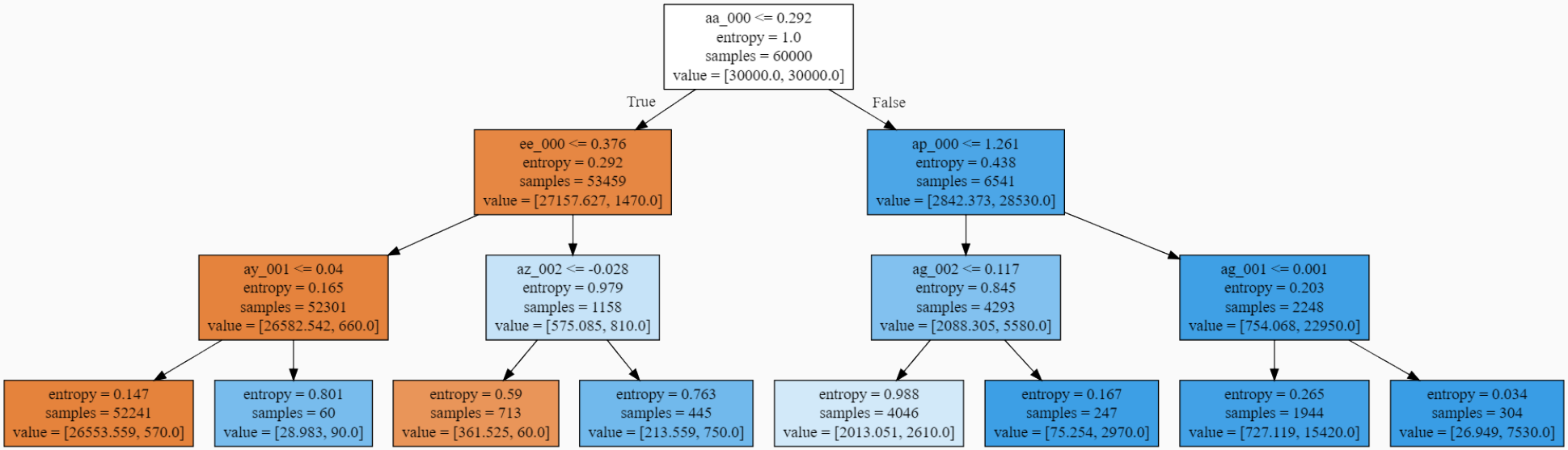
- *max_depth* : 1, 2, **3**, 4, 5, 6, 7, 8, 9, 10, 15, None
- *max_features* : **'sqrt'**, 'log2', 20, 30, 50, 100, 150, 170, None
- *min_samples_leaf* : **1**, 3, 5
- *criterion* : **'entropy'**, 'gini'

Results:

- *Recall* = 98.13%
- *Total Costs* = 19 010



Results: Decision Tree (cont.)



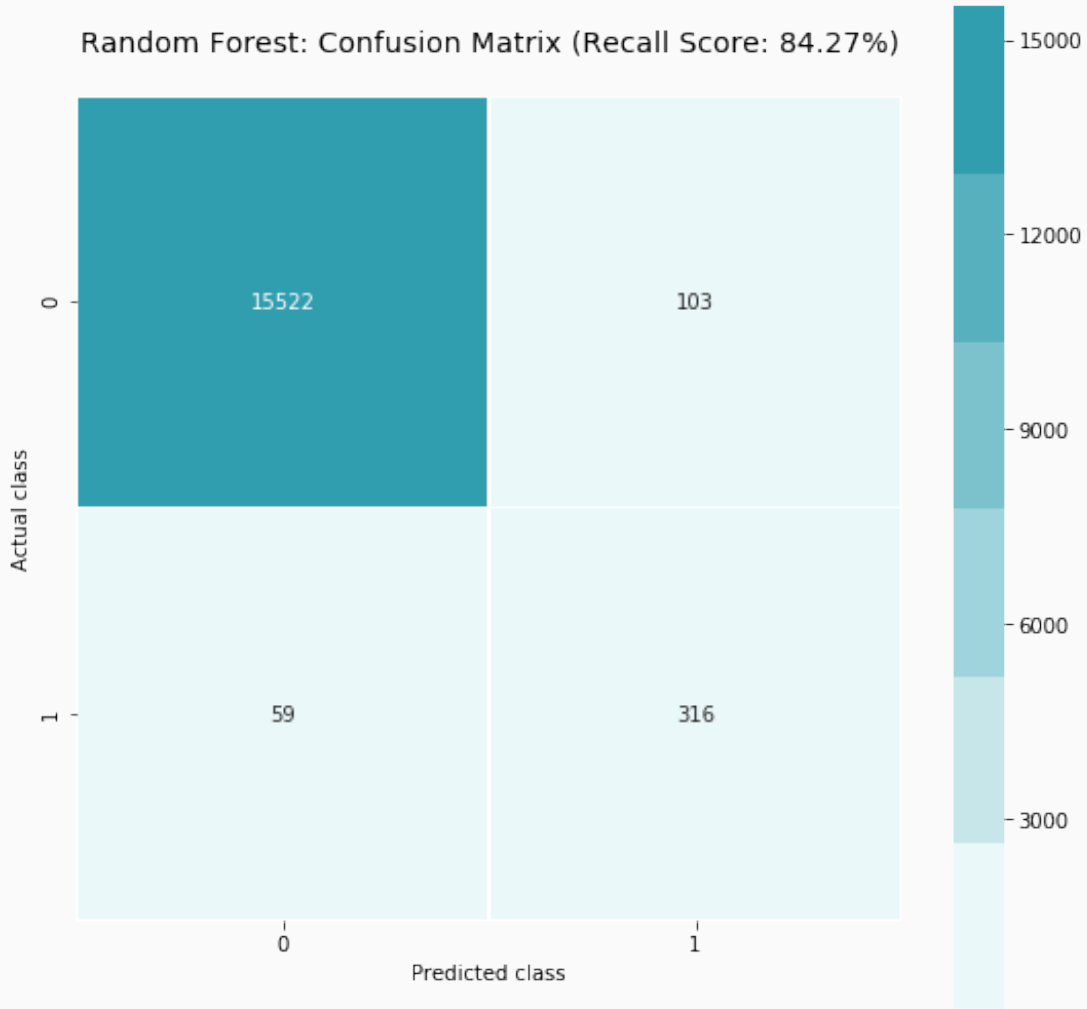
Results: Random Forest

Parameters:

- *n_estimators* : 100, 150, **200**
- *max_features* : 'sqrt', **'log2'**
- *min_samples_leaf* : 1, 3, **5**
- *max_depth* : **20**, 25, 30

Results:

- *Recall* = 84.27%
- *Total Costs* = 30 530



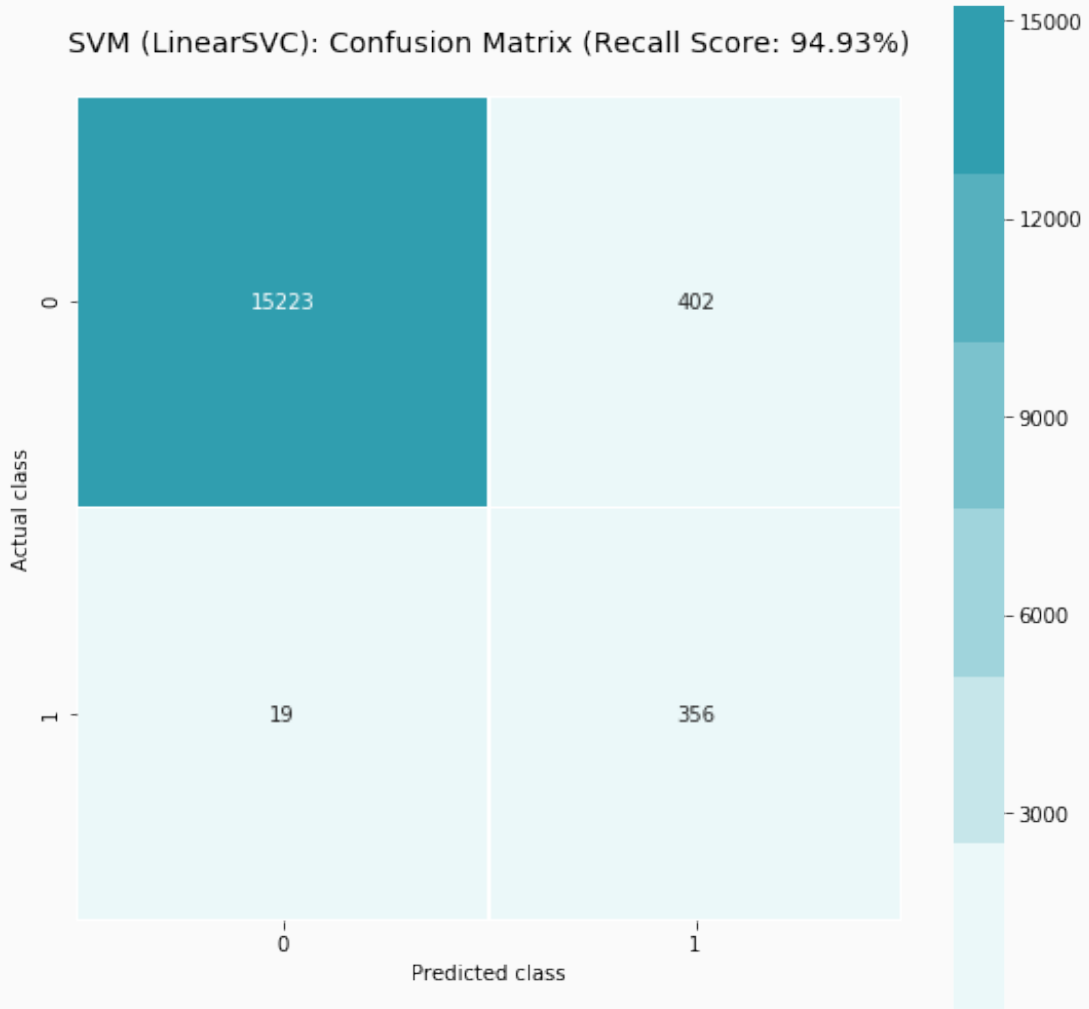
Results: SVM (LinearSVC)

Parameters:

- *C* : **0.0001**, 0.001, 0.01, 0.1, 1, 10, 100, 1000

Results:

- *Recall* = 94.93%
- *Total Costs* = 13 520



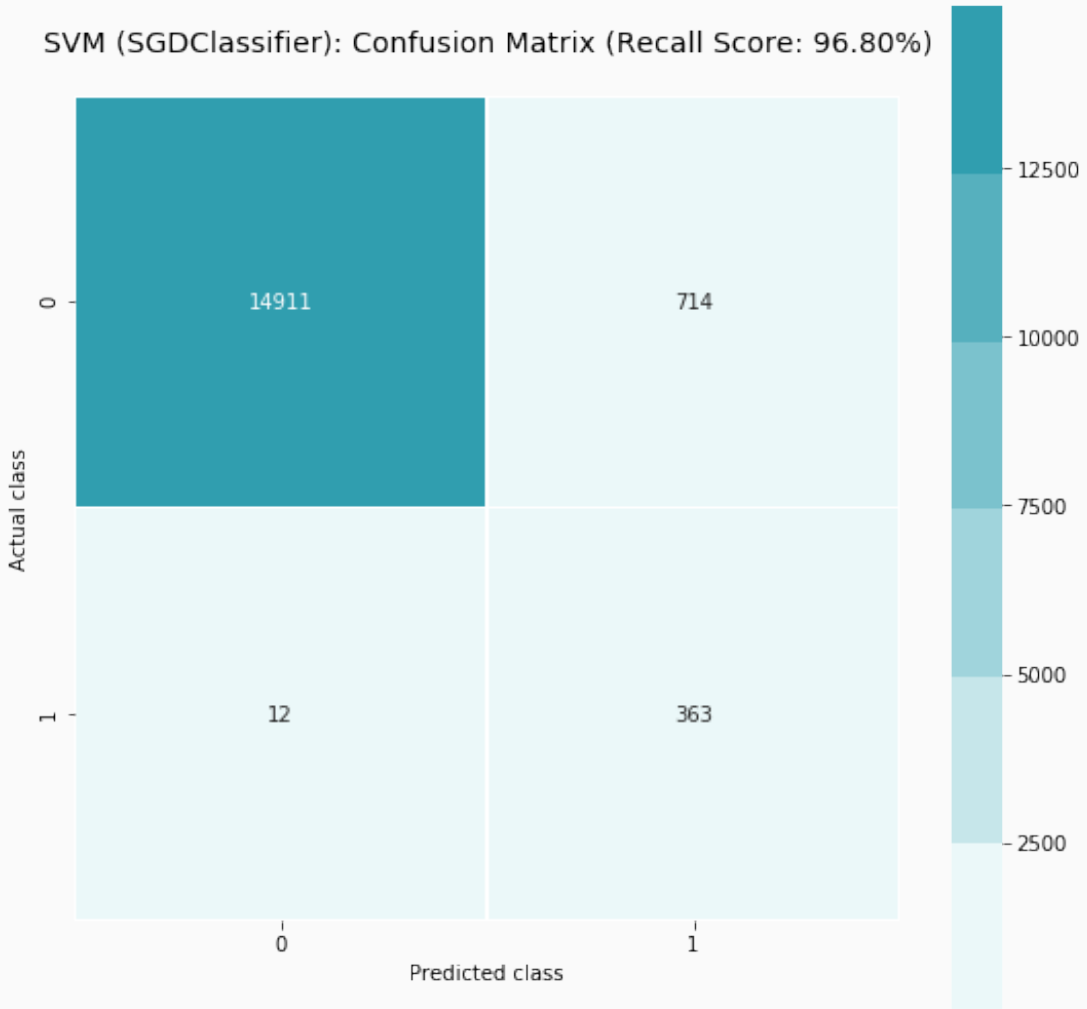
Results: SVM (SGDClassifier)

Parameters:

- *alpha* : **0.0001**, 0.001, 0.01, 0.1, 1, 10, 100, 1000

Results:

- *Recall* = 96.80%
- *Total Costs* = 13 140



Results: SVM (Non-linear)

Kernels:

- *RBF* (Radial Basis Function):

$$K = exp(-\gamma||x - x'||^2)$$

- *sigmoid*:

$$K = tanh(\gamma < x, x' >)$$

- *poly* (polynomial):

$$K = (\gamma < x, x' >)^3$$

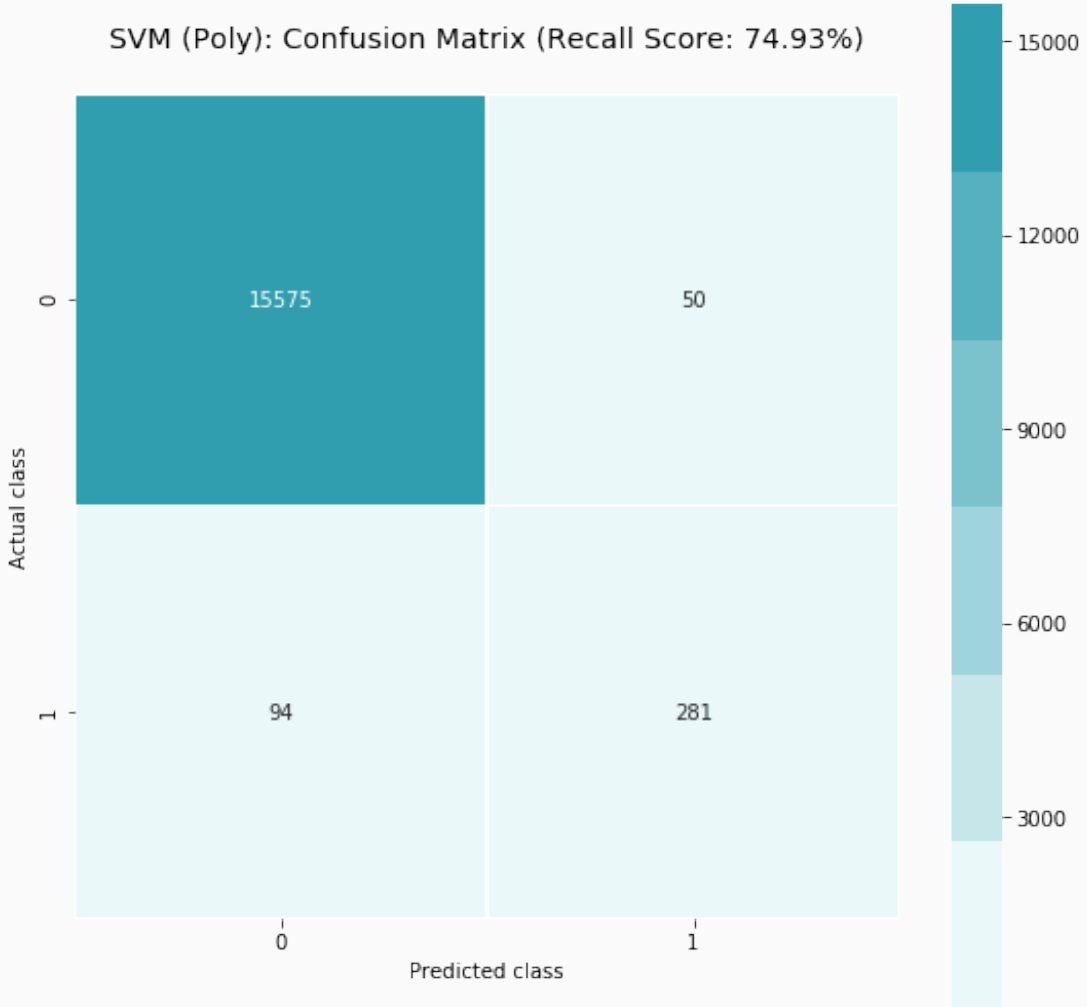
Results: SVM (Non-linear) (cont.)

Parameters:

- *kernel* : 'rbf', 'sigmoid', **'poly'**
- *gamma* : 0.0001, **0.001**
- *C* : 0.001, 0.01, 0.1, 1, 10, 100, **1000**

Results:

- *Recall* = 74.93%
- *Total Costs* = 47 500



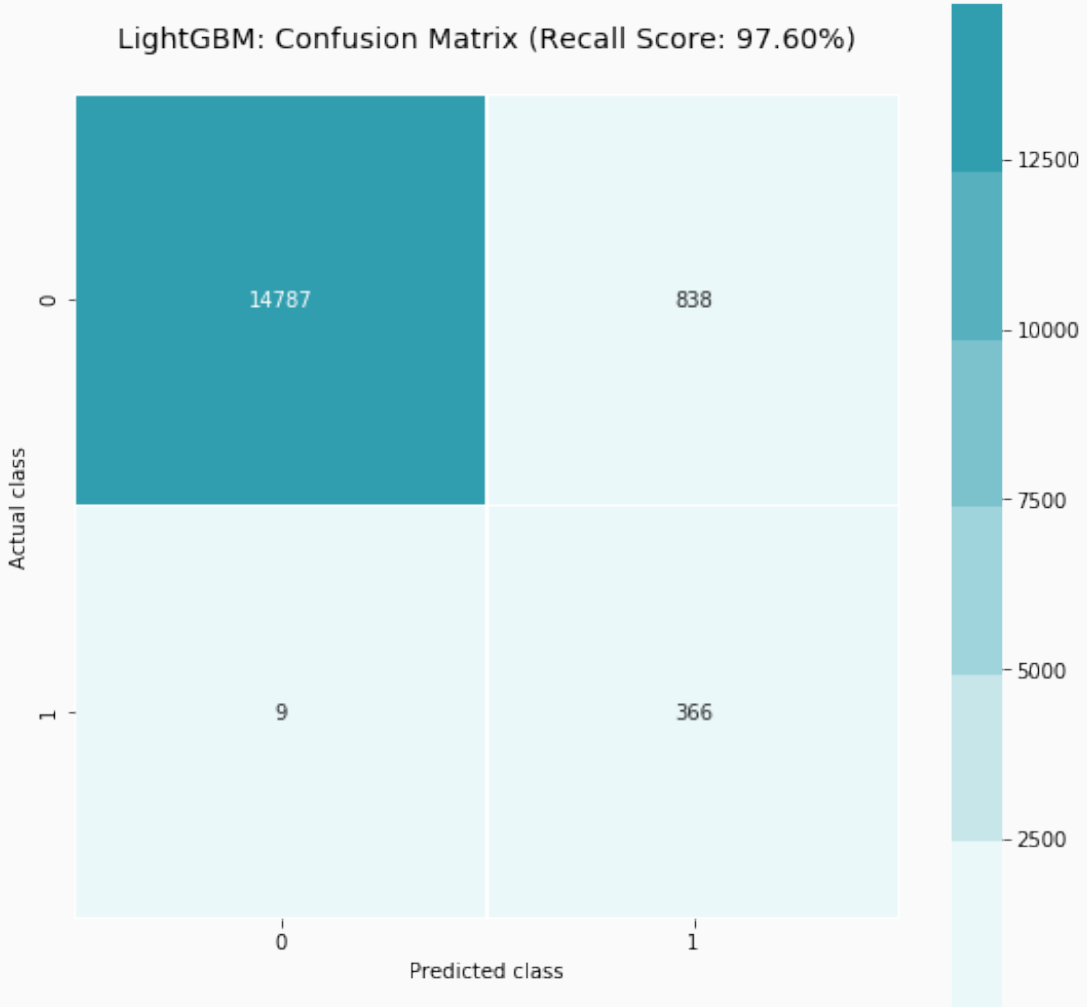
Results: LightGBM

Parameters:

- *learning_rate* : **0.001**, 0.005, 0.01, 0.1, 1
- *n_estimators* : **20**, 50, 100
- *num_leaves* : **10**, 20, 31

Results:

- *Recall* = 97.60%
- *Total Costs* = 12 880



Comparison of Results

Algorithm	Recall, %	Cost	Time, s	Comb	T/Comb, s
Logistic Regression	94.13	14 840	37.6	9	4.2
Decision Tree	**98.13**	19 010	2304	648	3.6
Random Forest	84.27	30 530	9228	108	85.4
SVM (Poly)	74.93	47 500	2136	42	50.9
SVM (LinearSVC)	94.93	13 520	444	8	55.5
SVM (SGDClassifier)	96.80	13 140	7.3	8	**0.9**
LightGBM	97.60	**12 880**	378	45	8.4

Possible Improvements

Pre-Processing:

- Removing outliers and features, that have low correlation with the target attribute

Feature Selection:

- PCA, Recursive Feature Elimination

Parameter Optimization:

- RandomizedSearch, HyperOpt

Scoring Function:

- Custom scoring function instead of Recall

Conclusion

Is Predictive Analysis worth it?

- Case 1 (”arrogant”):
  $\textit{Total Costs} = 375 \cdot 500 = 187\,500$
- Case 2 (”safe”):
  $\textit{Total Costs} = 16\,000 \cdot 10 = 160\,000$
- Predictive Analysis:
  $\textit{Total Costs} \approx 13\,000$

Thank you!