accenture

High performance. Delivered.

Case Study

Bankruptcy prediction

The process

Bankruptcy prediction

1 **Exploratory Data Analysis**
- Distributions
- Correlation

2 **Data Cleaning**
- Feature imputation
- Data generation
- Feature selection

3 **Training**
- Data split
- Model selection
- Optimization

4 **Results**
- Accuracy
- Confusion matrix

5 **Serving**
- Model serving techniques

The data

Bankruptcy prediction

1

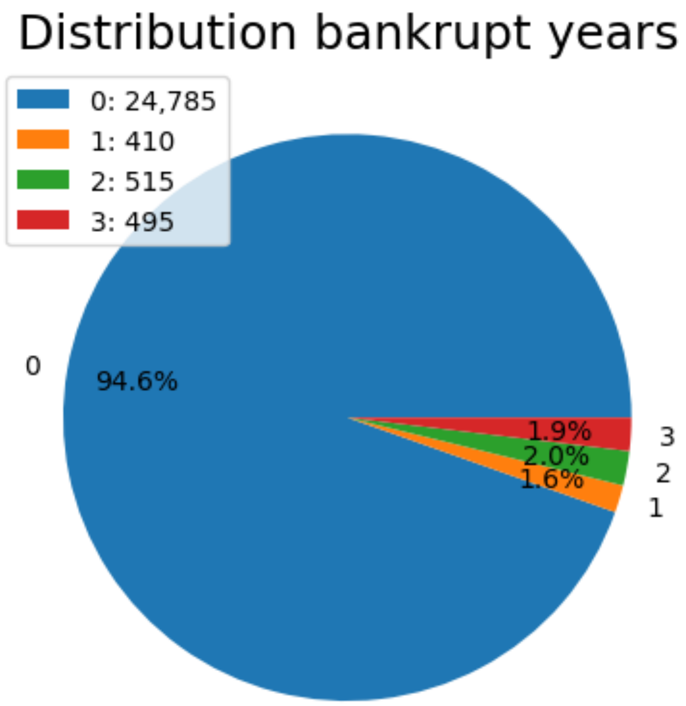
Year of documentation	Total companies	Not bankrupt after 5 years	Bankrupt after 5 years	Prediction Horizon	Number features
1 year	7027	6756	271 (3,86%)	5 years	65
2 year	10173	9773	400 (3,93%)	4 years	65
3 year	10503	10008	495 (4,37%)	3 years	65
4 year	9792	9277	515 (5,26%)	2 years	65
5 year	5910	5500	410 (6,94%)	1 year	65

- We want to predict the probability for a company to go bankrupt after 1, 2 or 3 years.
  - Documentation horizon is from year 3 to year 5
  - For each class we need to predict a probability
    - [p(notB), p(1year), p(2years), p(3years)]
- Data set is imbalanced

The training subset

Bankruptcy prediction

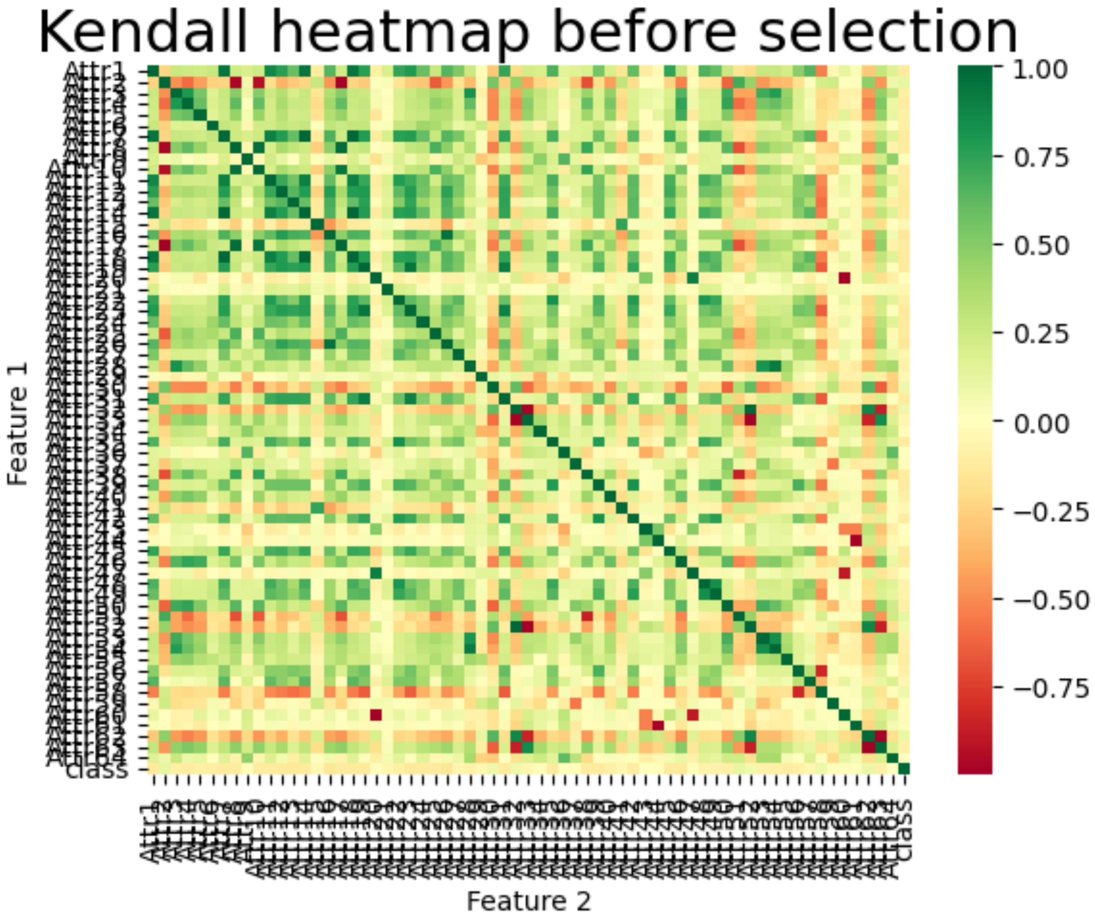
- For each company enitity we assign a label
  - 0 if did not go bankrupt
  - 1 if went bankrupt after 1 year
  - 2 if went bankrupt after 2 years
  - 3 if went bankrupt after 3 years
- We then concatenate the tables to one
- The features available contained nil values
  - We imputed them with the mean



Feature selection

Bankruptcy prediction

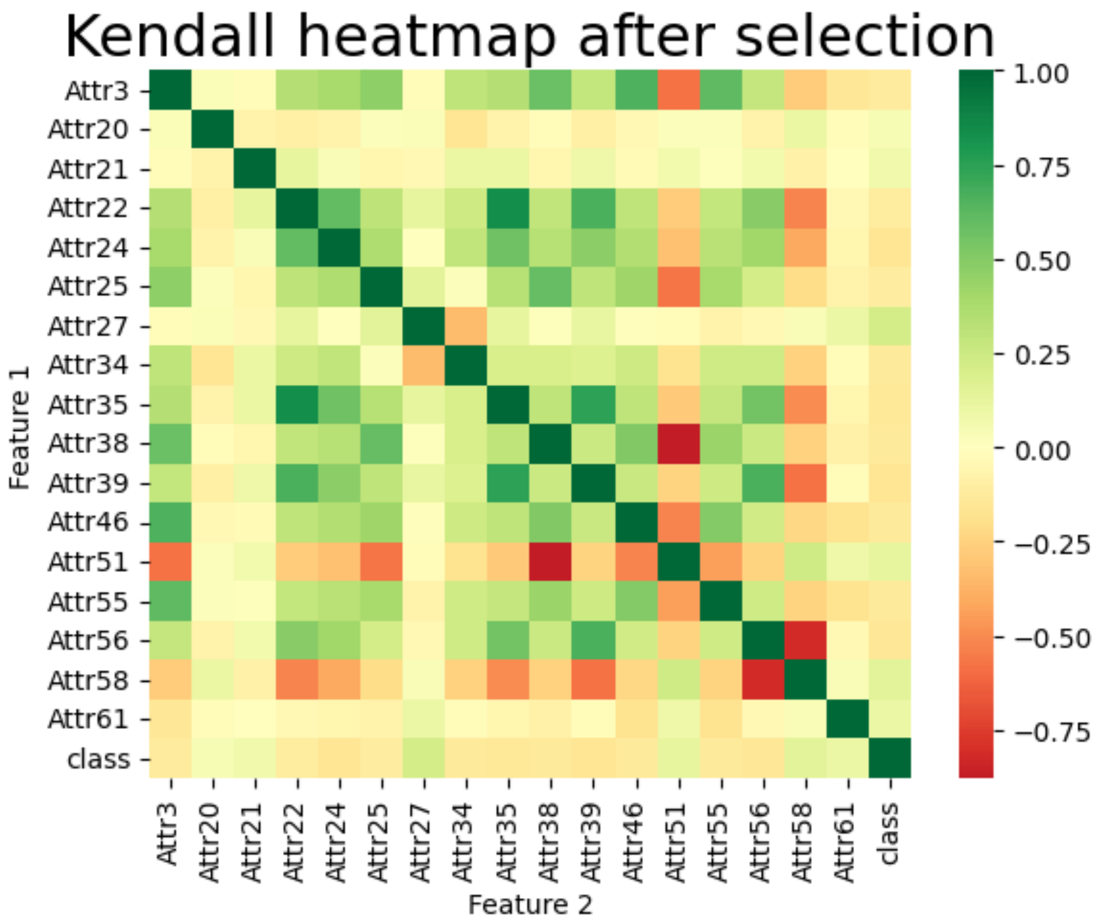
- Many similar features which have the same or almost equal Information
- Keeping both would falsify the model
- Therefore we need to select the features with the highest feature importance for our model



Feature selection

Bankruptcy prediction

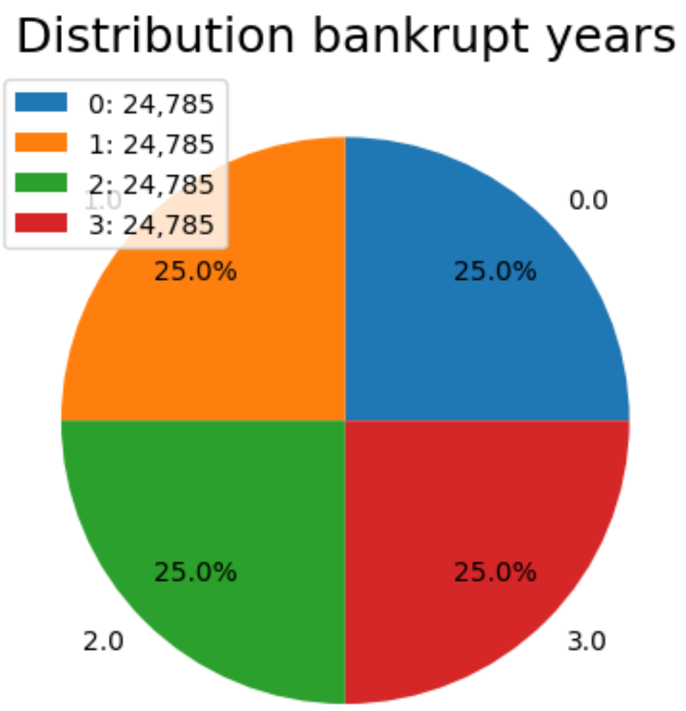
- With an ensemble technique we calculate for each feature in the dataset a feature importance score
- To select the right features we compared the
  - ExtraTreesClassifier and the
  - SelectKBestClassifier
- We later used the ExtraTreesClassifier
- The most important features are selected reducing the number of features from 65 to 18



Fixing the imbalance

Bankruptcy prediction

- We use a **S**ynthetic **M**inority **O**versampling **Te**chnique to adress the imbalance of the dataset
- As our classes are close to each other and SMOTE uses a Kneighbor Model we used borderlineSMOTE
- That means, that we create new entities for the under represented classes
- Now we have 99,140 entities



Model selection

Bankruptcy prediction

3

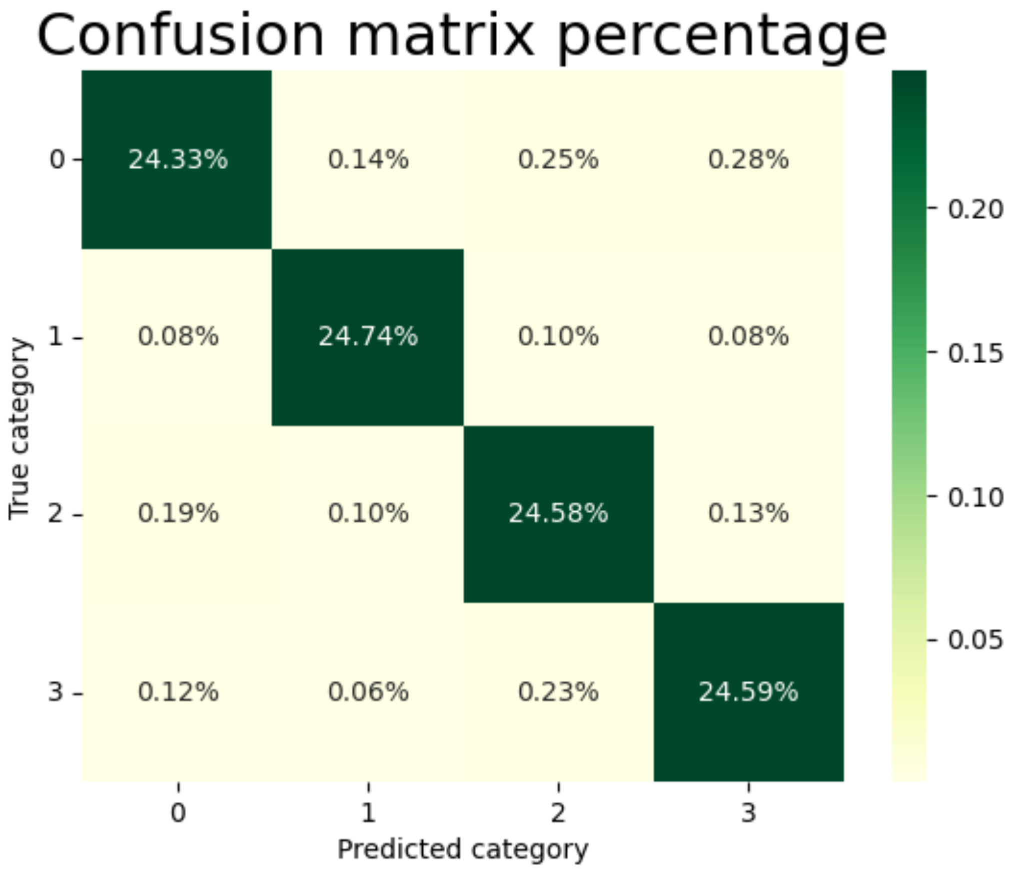
- For our multi classification task we tested different models
- Number of entities: 99,140 (80/20 split)
- Number of features: 18
- Closer distinction between XGBoost and RandomForest
- We use XGBoost due to lower random influence than RF

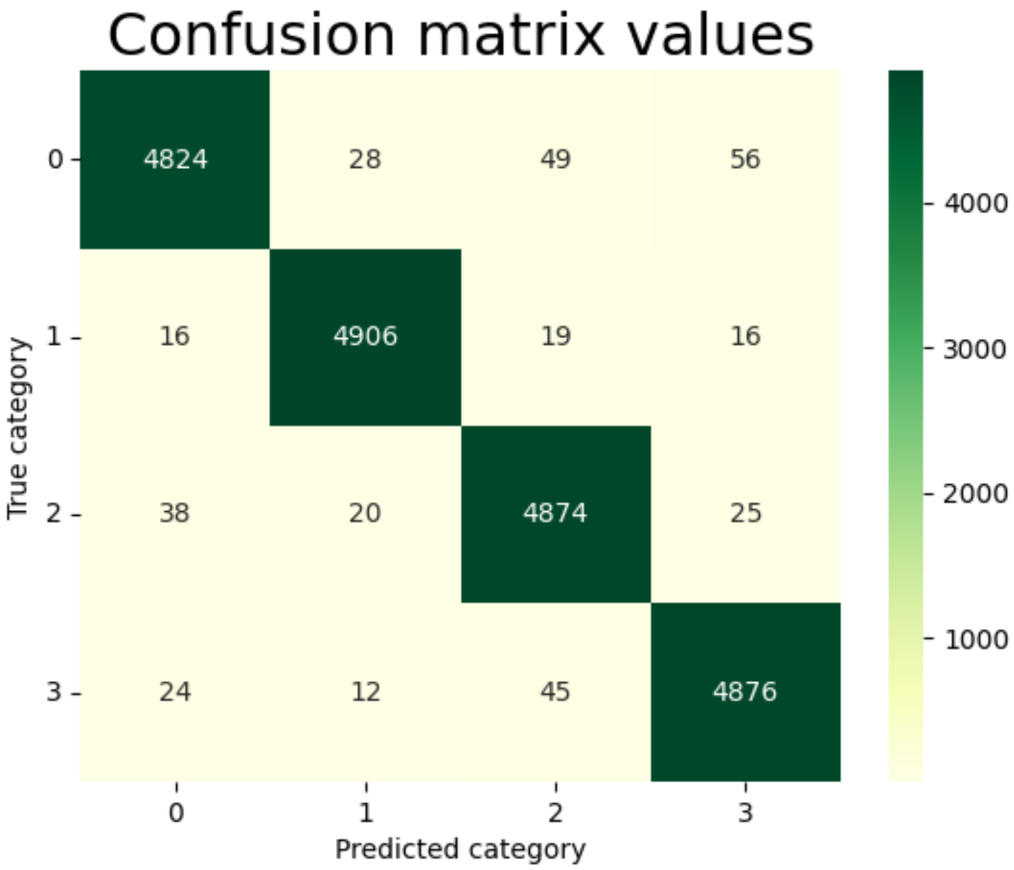
Model	Accuracy
DecisionTree	92.59
RandomForest	98.61
AdaBoost	50.51
KNeighbors	77.87
XGBoost	98.24

Results

Bankruptcy prediction

Accuracy: 98,24490619326205





Model serving

Bankruptcy prediction

5

To include the insights in a dashboard used by their risk managers we need to:

1. Setup a webserver which will function as a host for a REST API
2. The to be interpreted data needs to be fit to the shape of our model
3. With a REST request, we can retrieve the values of our model and build the insights for the risk manager

- Our implementation is using pythons ‘mlserver’ to host a mock of the REST API locally

Case Study
Bankruptcy prediction

Thank you!

High performance. Delivered.

Model results detailed

Backup slide

	Precision	Recall	F1-score	Support
0.0	0.9841	0.9732	0.9786	4957
1.0	0.9879	0.9897	0.9888	4957
2.0	0.9773	0.9833	0.9803	4957
3.0	0.9805	0.9837	0.9821	4957