Machine Learning Projects

8th October 2018

Classification problem

Delicatessen Company

Goal

- Prediction of the next marketing campaign (binary classification)
- Most profit possible by contacting the customers that have the higher probability of accepting the offer.
- ~15% response rate
- Restriction: 9 features
- Per client:
  - If accepts the offer → Profit: 11
  - If rejects the offer → Cost: 3

Dataset

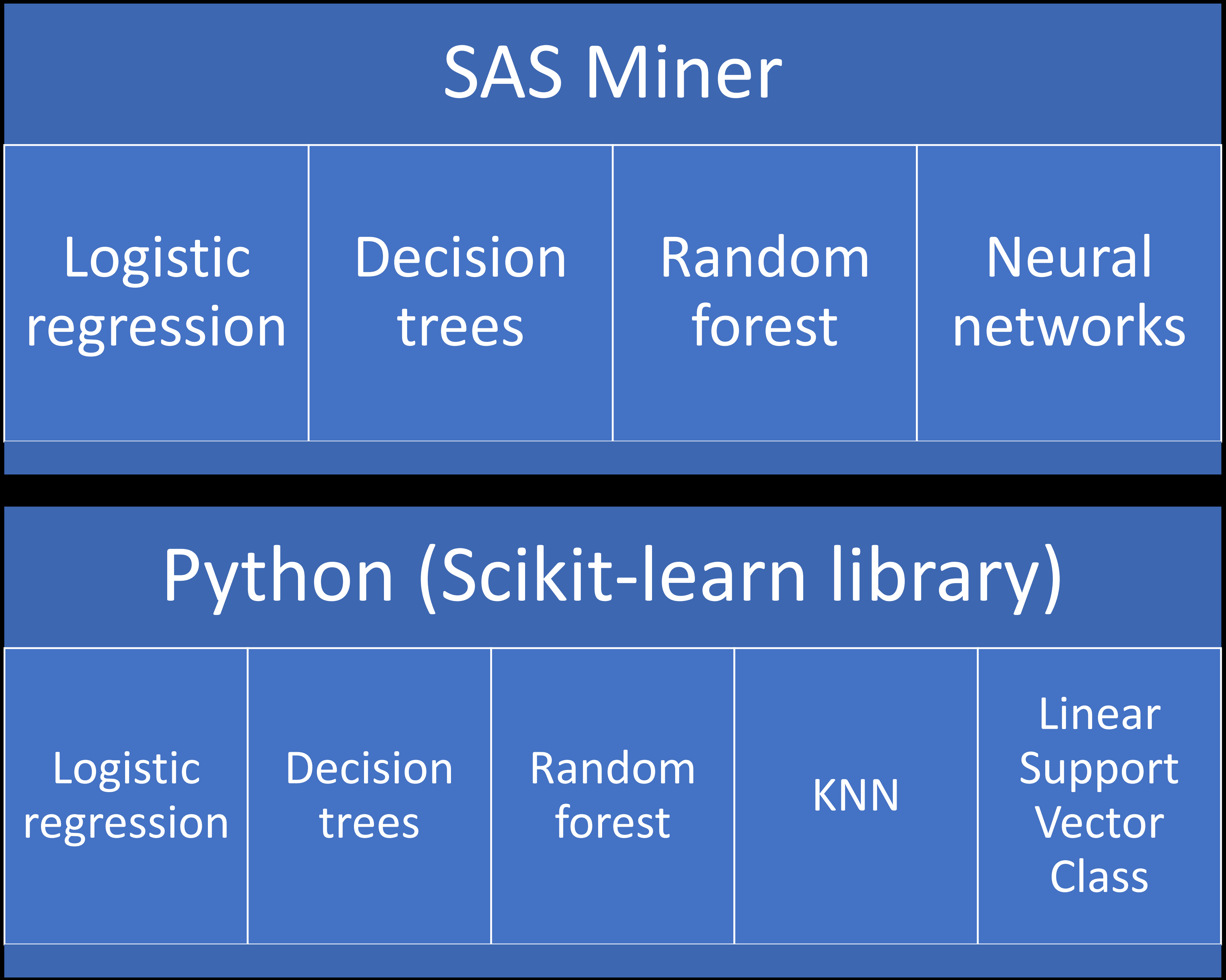
Build a classifier:

- 1450 customers
- 30 predictors and 1 target variable (‘DepVar’).
- Data Split:
  - Train set: 70% ~ 1015 customers
  - Test set: 30% ~ 435 customers
  - Cross-validation: 5

Predict target class:

- 1855 new customers
  - Subset 15%/20% of the customers to be contacted for the next campaign (Teacher had the target values (0 or 1) and therefore knew the real profit)

Models



SAS Miner

- Logistic regression
- Decision trees
- Random forest
- Neural networks

Python (Scikit-learn library)

- Logistic regression
- Decision trees
- Random forest
- KNN
- Linear Support Vector Class

Feature Selection

- PCA
- Removing highly correlated variables – Redundancy
- Variable importance: Decision tree and random forest
- Regression selection models (forward, backward, stepwise)

Model comparison

- Test score: metric ROC/AUC

Best model:

- SAS:
  - Neural network with 20 hidden neurons, 0 weight decay and loss function: misclassification
  - Real profit ~900

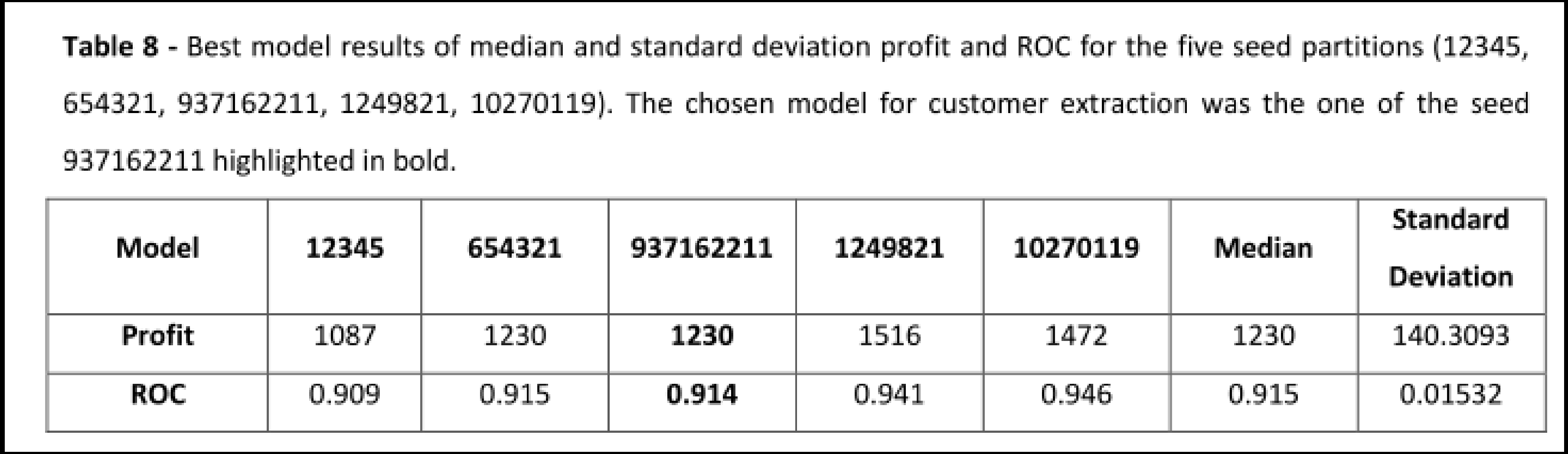


Table 8 - Best model results of median and standard deviation profit and ROC for the five seed partitions (12345, 654321, 937162211, 1249821, 10270119). The chosen model for customer extraction was the one of the seed 937162211 highlighted in bold.

Model	12345	654321	937162211	1249821	10270119	Median	Standard Deviation
Profit	1087	1230	**1230**	1516	1472	1230	140.3093
ROC	0.909	0.915	**0.914**	0.941	0.946	0.915	0.01532

Regression problem

Human oral bioavailability of a new drug

Goal

- Prediction of human oral bioavailability of a new drug as a function of its molecular descriptors (%F) using genetic programming.
- Names of the columns were unknown and was a concatenation of different data sets:
  - prediction of the Unified Parkinson's Disease Rating Scale (19 features);
  - prediction of high performance concrete strength (9 features);
  - prediction returns of the Istanbul Stock Exchange (8 features);
  - predicting the value of human oral bioavailability (242 features).
- The first three data sets are, for obvious reasons, problem-unrelated.

Dataset

- 277 unknown features + 1 continuous target variable
- 282 instances
- Feature selection:
  - Remove columns only containing zeros (no variability)
- Split data:
  - Training set: 70%
  - Test set: 30%

Genetic Programming model (Java)

- Hyperparameter tuning:
  - Different crossovers
  - Mutations
  - Selections
  - Elitism
- Restrictions:
  - Population size: 300
  - Number of generations: 300

Best individual of the population
Tree-based model

Model comparison (cv: 10):
Metric: Root Mean Squared Error
Best training error
Best unseen error (test set)
Best absolute error (best unseen error with the best training error)

Genetic Programming model (Java)

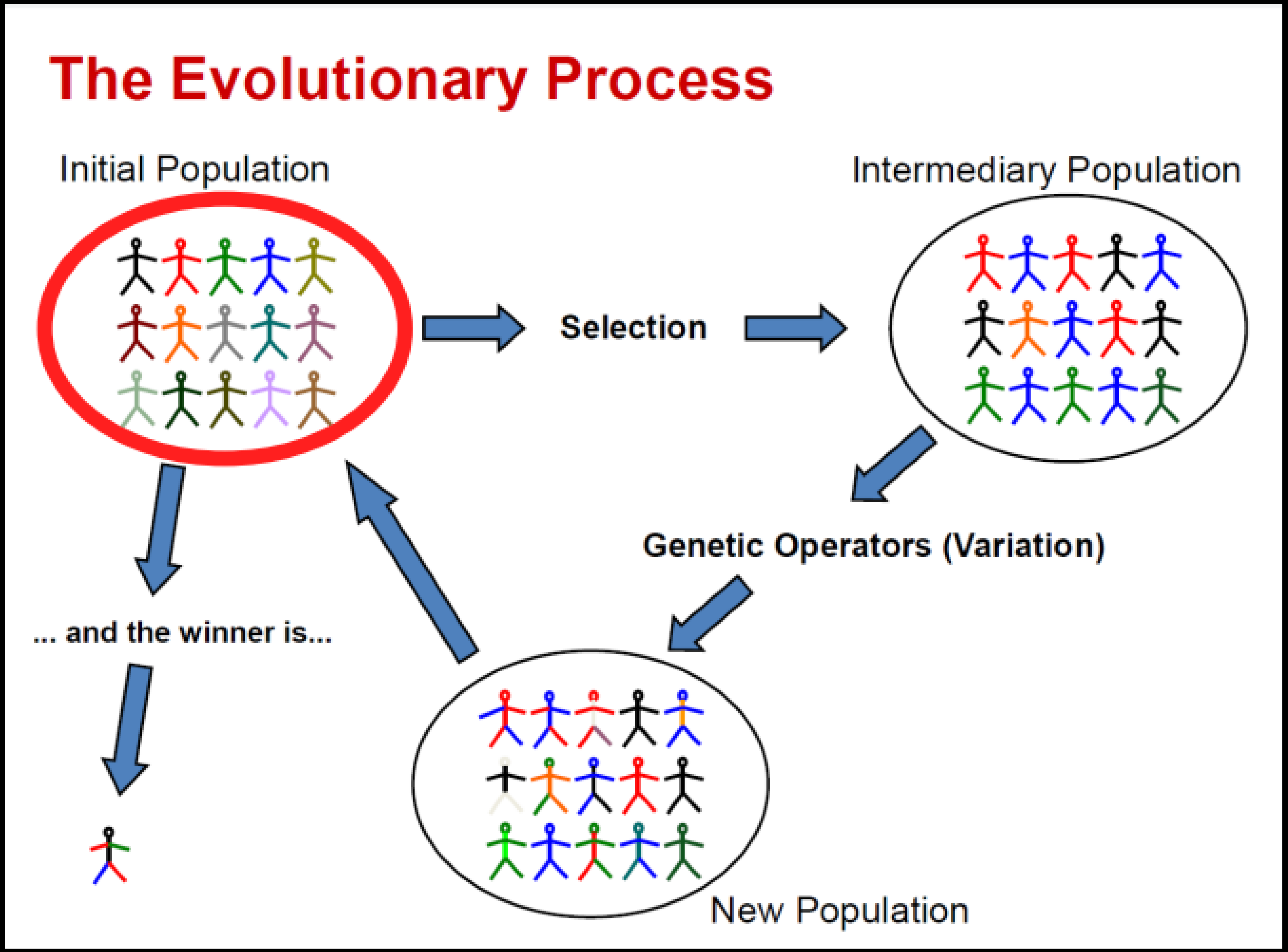
- Best parameters:
    - Parallel population
        - Number of pop: 8
    - Mutation
        - Type: default
        - Number: 1
    - Elitism
        - 100 best individuals
    - Selection
        - Type: Tournament
        - Pressure: 20 individuals

- Crossover
    - Type: Default
    - Prob: 0.3

Best model

- Final individual:
    - Training error: 23.73
    - Test error: 26.67

    - Tree size: 490
    - Depth: 35


The Evolutionary Process
Initial Population
Selection
Intermediary Population
Genetic Operators (Variation)
... and the winner is...
New Population