CAR INSURANCE FRAUD DETECTION

Machine Learning Lab Project

AGENDA

Business Understanding

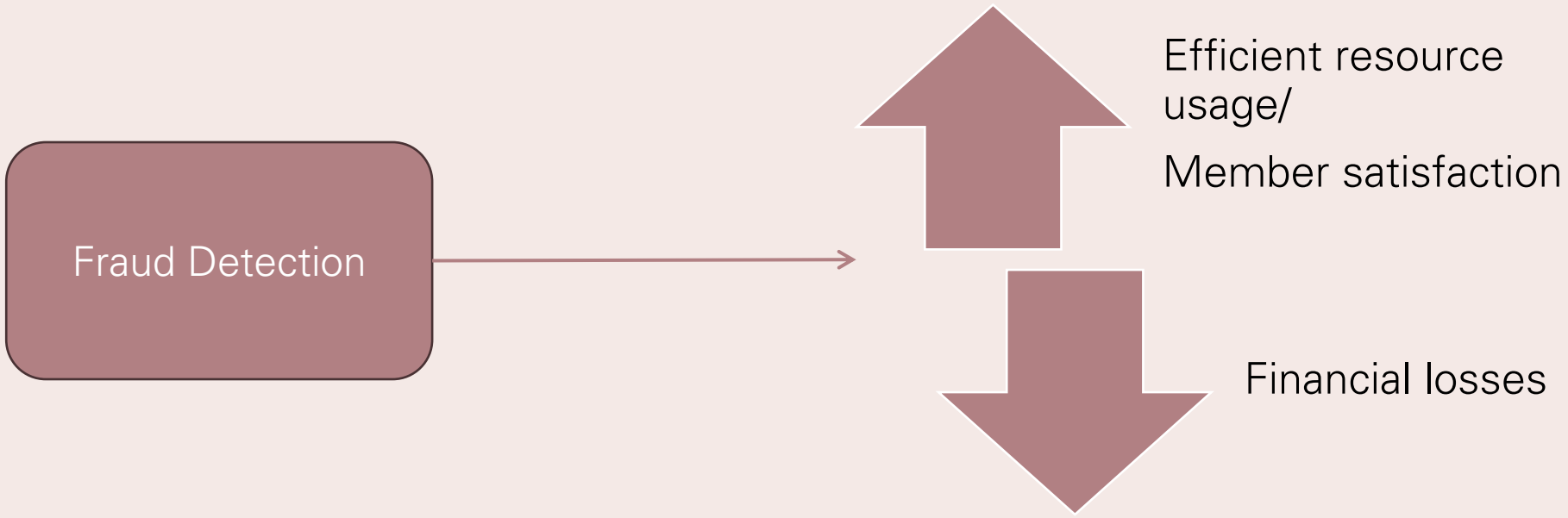
Data Understanding

Data Preparation

Modeling

Evaluation

BUSINESS UNDERSTANDING



DATA UNDERSTANDING

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   policy_id                    30000 non-null  object
 1   policy_state                 30000 non-null  object
 2   policy_deductible            30000 non-null  int64
 3   policy_annual_premium        30000 non-null  float64
 4   insured_age                  30000 non-null  int64
 5   insured_sex                  30000 non-null  object
 6   insured_education_level      30000 non-null  object
 7   insured_occupation           30000 non-null  object
 8   insured_hobbies              30000 non-null  object
 9   incident_date                30000 non-null  object
 10  incident_type                30000 non-null  object
 11  collision_type               30000 non-null  object
 12  incident_severity            30000 non-null  object
 13  authorities_contacted        22436 non-null  object
 14  incident_state               30000 non-null  object
 15  incident_city                30000 non-null  object
 16  incident_hour_of_the_day     30000 non-null  int64
 17  number_of_vehicles_involved  30000 non-null  int64
 18  bodily_injuries              30000 non-null  int64
 19  witnesses                    30000 non-null  int64
 20  police_report_available      30000 non-null  object
 21  claim_amount                 30000 non-null  float64
 22  total_claim_amount           30000 non-null  float64
 23  fraud_reported               30000 non-null  object
dtypes: float64(3), int64(6), object(15)
```

DATA UNDERSTANDING

- Synthetic dataset
- Missing values
- No major outliers
- Many categorical data -> have to be converted into numerical data
- Target variable is highly imbalanced

DATA PREPARATION

Numerical Data

1. Drop unnecessary columns
2. One-hot-encoding
3. Label encoding
4. Target encoding
5. Normalization
6. Balancing

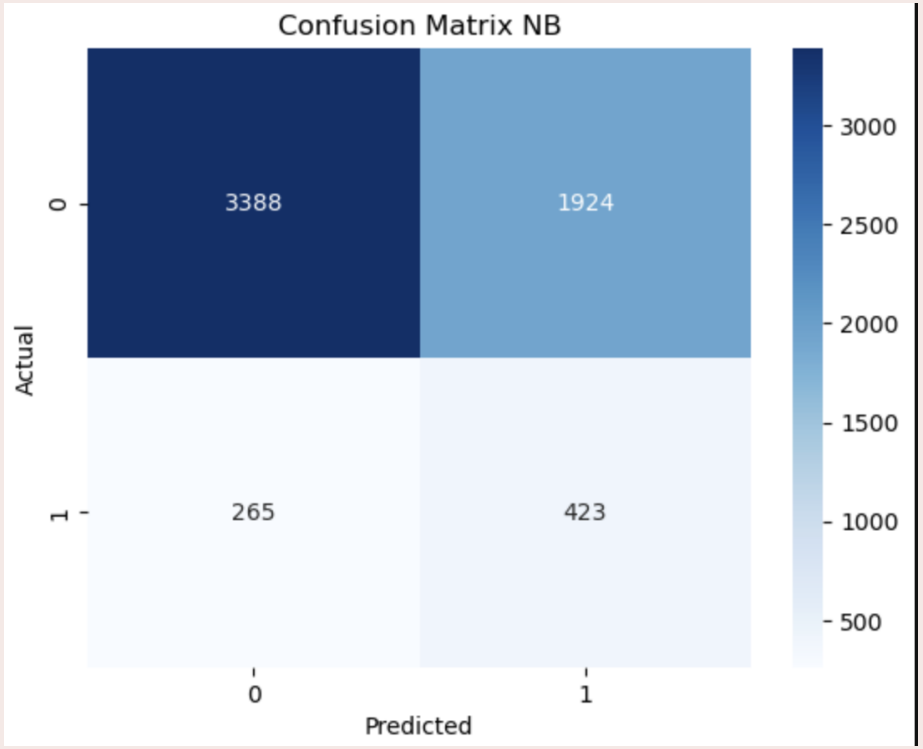
Categorical Data

1. Drop unnecessary columns
2. Ordinal encoder
3. Binning
4. Balancing

MODELING

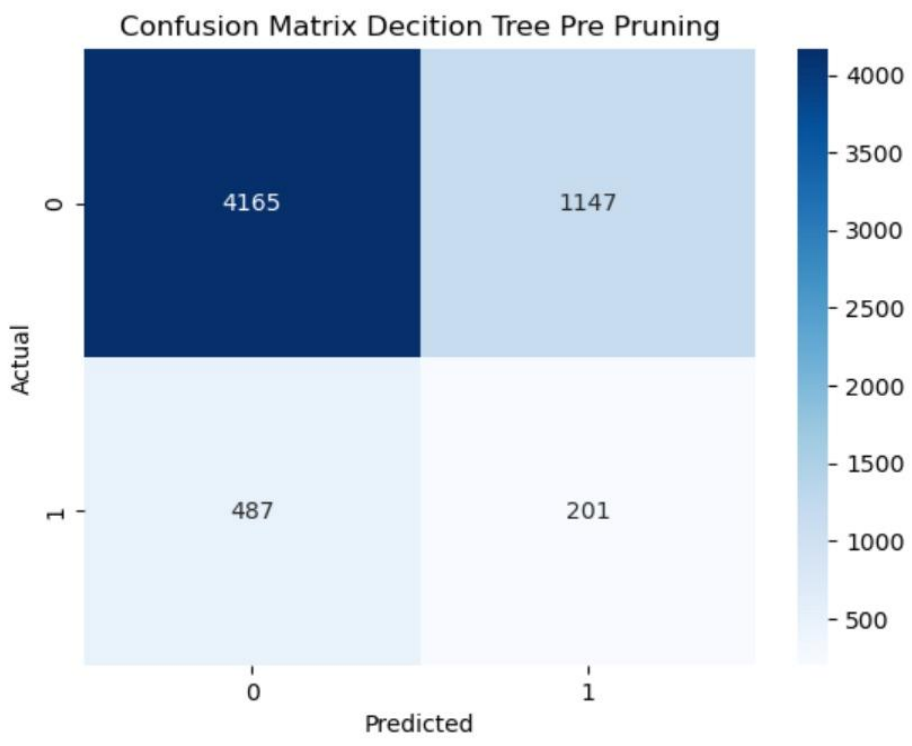
- Naive Bayes
- Decision Tree
- Random Forest
- k-NN
- ANN

CATEGORICAL NAIVE BAYES



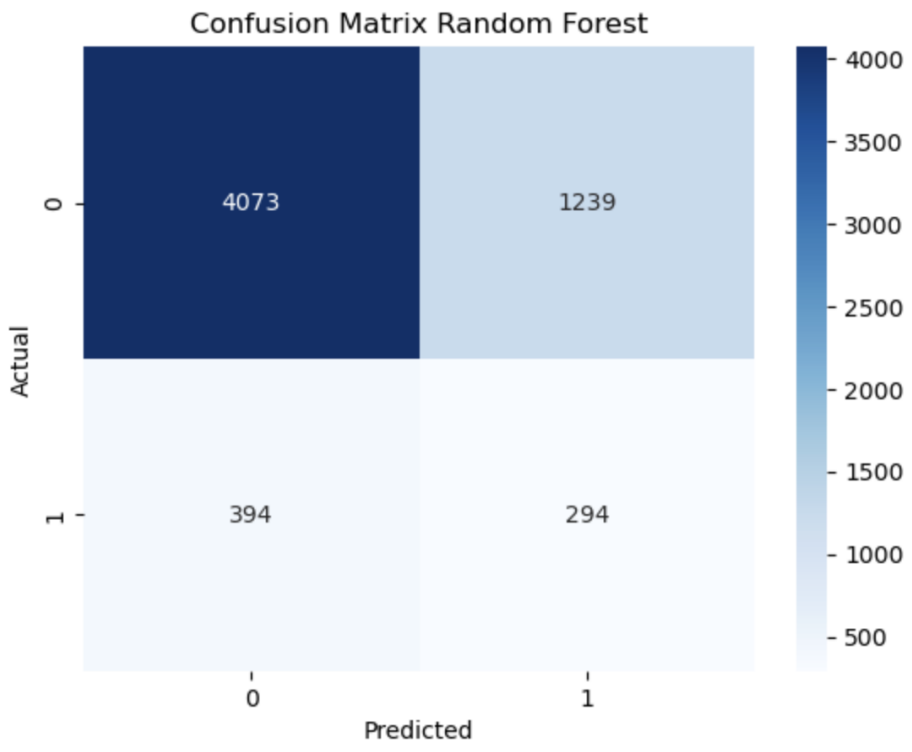
	precision	recall	f1-score	support
N	0.93	0.64	0.76	5312
Y	0.18	0.61	0.28	688
accuracy			0.64	6000
macro avg	0.55	0.63	0.52	6000
weighted avg	0.84	0.64	0.70	6000

DECISION TREE WITH POST-PRUNING



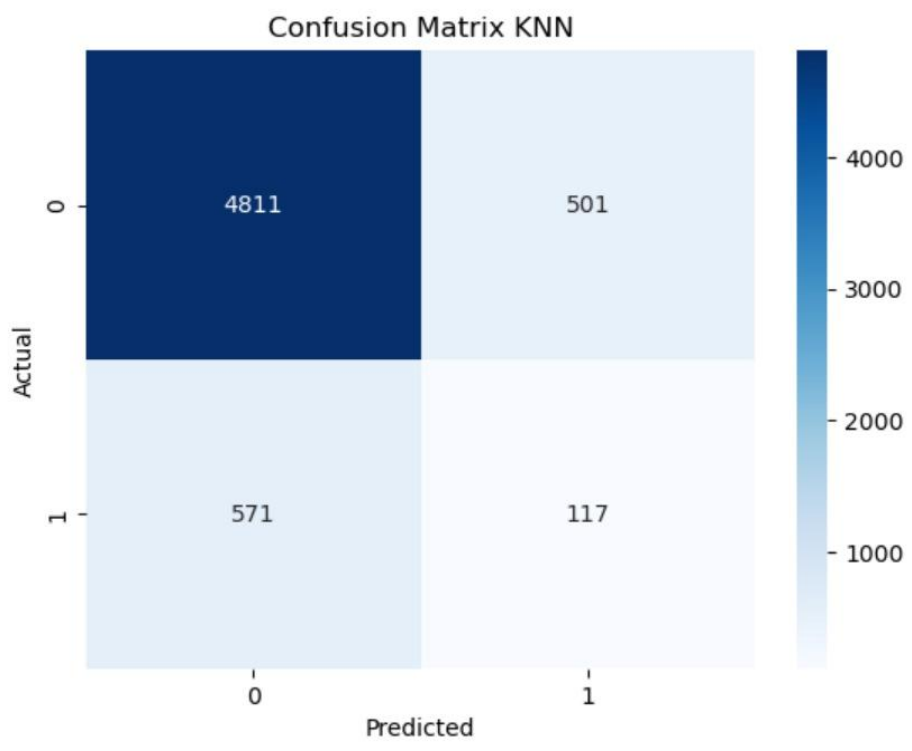
	precision	recall	f1-score	support
0	0.90	0.78	0.84	5312
1	0.15	0.29	0.20	688
accuracy			0.73	6000
macro avg	0.52	0.54	0.52	6000
weighted avg	0.81	0.73	0.76	6000

RANDOM FOREST



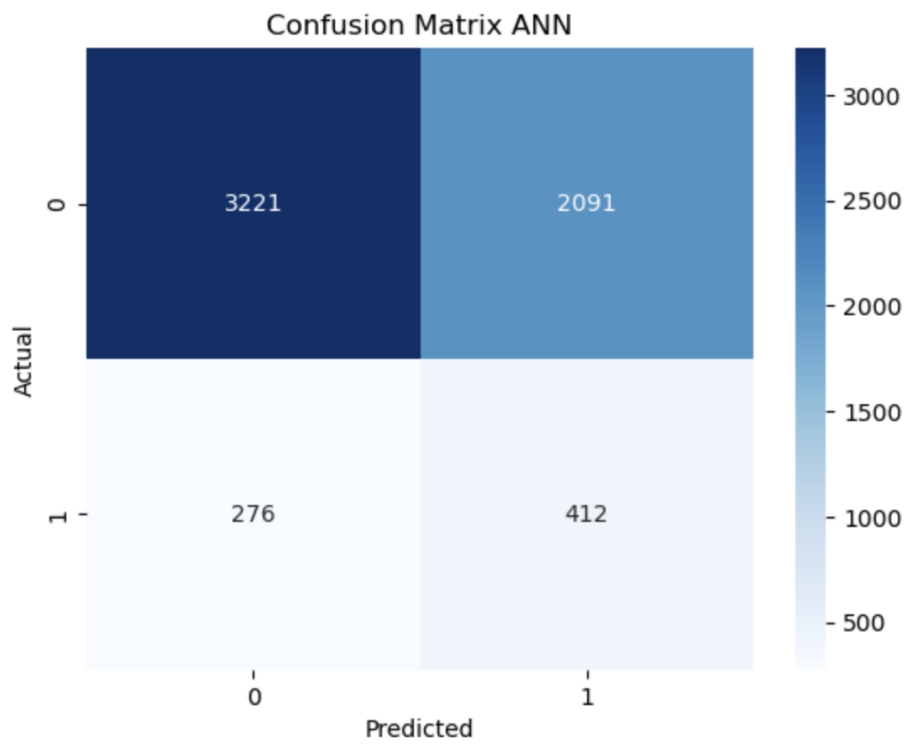
	precision	recall	f1-score	support
0	0.91	0.77	0.83	5312
1	0.19	0.43	0.26	688
accuracy			0.73	6000
macro avg	0.55	0.60	0.55	6000
weighted avg	0.83	0.73	0.77	6000

K-NN



	precision	recall	f1-score	support
0	0.89	0.91	0.90	5312
1	0.19	0.17	0.18	688
accuracy			0.82	6000
macro avg	0.54	0.54	0.54	6000
weighted avg	0.81	0.82	0.82	6000

ARTIFICIAL NEURAL NETWORK



	precision	recall	f1-score	support
0	0.92	0.61	0.73	5312
1	0.16	0.60	0.26	688
accuracy			0.61	6000
macro avg	0.54	0.60	0.49	6000
weighted avg	0.83	0.61	0.68	6000

EVALUATION

- Comparison on accuracy not possible
- F1 score, recall and precision more suitable
- Many false positives across all models
- Fixing leads to decrease in fraud detection in general
- F1 score of max. 0.28 across all models
- Models not optimal for deployment
- -> Business goal not achieved