ML in Insurance

Oliver • Mads • Niklas • Kristoffer • Jacob

Final Project - Applied Machine Learning 2026

Domain Specific Motivation

How can insurance firms utilize Machine learning models when pricing insurance premiums from claims data, measuring the risk of insurance fraud from policy information and detecting insurance fraud when handling claims?

Pricing

- Classical GLM model
- NN and BDT model
- Clustering model finding the average probability of fraud in each cluster
- Some way of taking the probability of fraud into account

Claims handling

- Classification model, what is the probability of fraud in this specific case.
- Image fraud detection, classification of fraud by image
- Utilize this to target your fraud investigation

Our Datasets

	Oracle fraud dataset	FreMLP CAS dataset
Type	Fraud data in comprehensive car insurance	Claims data in comprehensive car insurance
Rows	15,420	20,600
Number of variables	33	22
Catagorical variables	15	12
Continuos variable	5	6
Ordinal variables	13	4
Fraud/Claims	923 cases of fraud (5.99%)~6%)	844 claims reported (4.10%)~4%)
Data quality fixes	• Dropped 1 row with faulty information (*DayOfWeekClaimed=0, MonthClaimed = 0)*; • Imputed 320 *Age=0* rows with corresponding *AgeOfPolicyHolder (Age 18)*	
Sample of variables	DayOfWeek, Accident area, Age, DriverRating, AgeOfVehicle, WitnessPresent, FraudFound_P	Exposure, Gender, VehAge, ClaimAmount, DrivAge

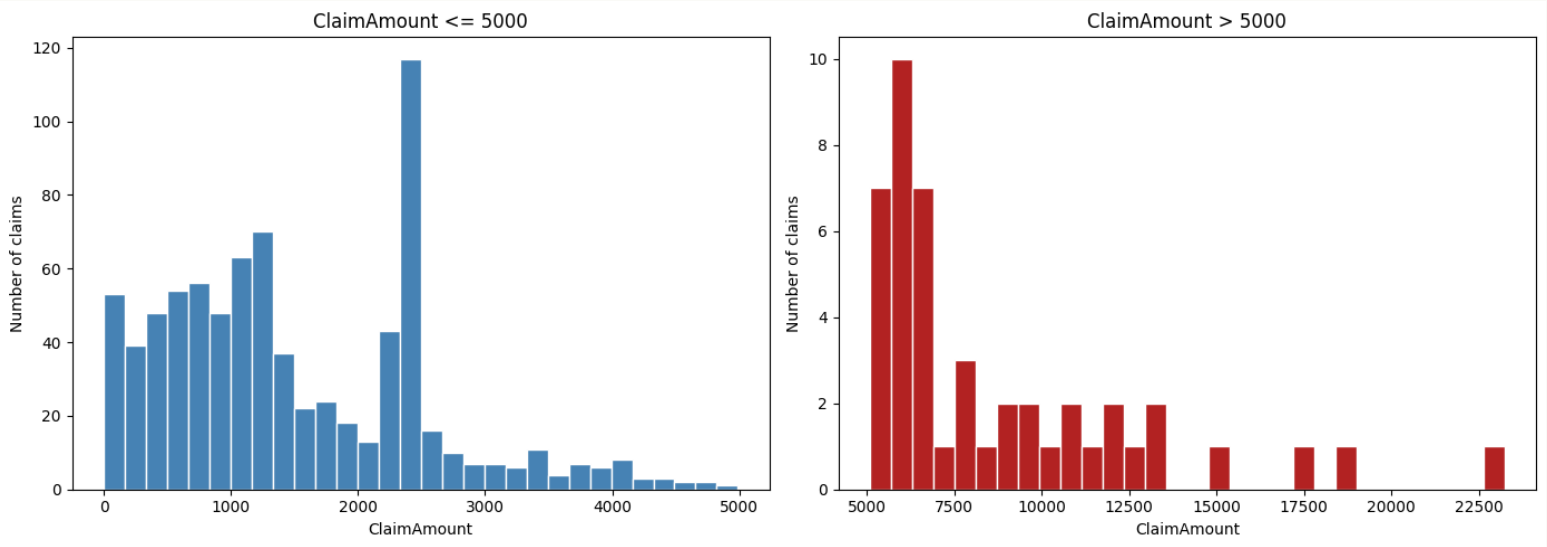
Tariff baseline — Tweedie GLM

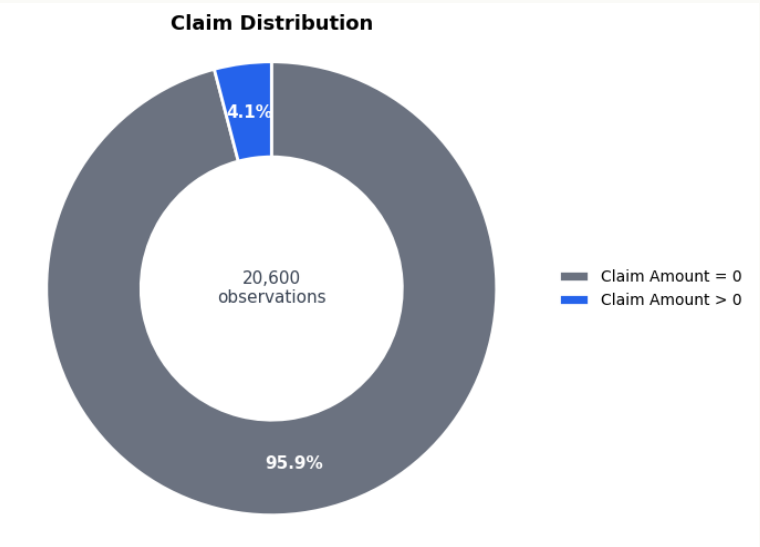
- **Method:** Tweedie GLM on claims below 5.000€ + large claim surcharge using EVT methods (Classical actuarial approach)
- **Advantages:** High interpretability and transparency, easy to manipulate and easy to implement
- **Disadvantages**: Limited flexibility (linearity assumptions), manual feature engineering, weak at complex interactions and high need of grouping which results in loss of information
- **Goal**: "Beat" the GLM Tweedie deviance using ML methods without compromising the normalized Gini coefficient and bias ratio

	Person 1	**Person 2**	**Person 3**
Gender	Male	Male	Female
Marital Status	Other	Alone	Other
Driver Age	56	26	51
Driver Rating	51	100	147
Vehicle Age	10+	3	0
Vehicle Price	J (Medium-Low)	M (Medium)	H (Medium-Low)
Risk Premium	168.59€	304.58€	235.50€

Initial choices

- Full data rather than Severity/Frequency
- Include large claims even though it might introduce bias
- All models are trained, tuned and evaluated on the same train/validation/test split
- Policy exposure must be included as an offset in each model
- License Age and Vehicle Engine are removed due to high correlation with Driver Age and Vehicle Energy





The XGBoost journey

$$\sum_i \text{TweedieLoss}(y_i, \mu_i) \quad \Rightarrow \quad \sum_i w_i \cdot \text{TweedieLoss}(y_i, \mu_i)$$

	Tweedie Deviance	Normalized Gini	Bias Ratio
Tweedie GLM	80.6596	0.1262	1.1978
Raw XGBoost	92.4066	0.0574	1.1943
Grouped XGBoost	82.6656	0.1498	1.1840
2x Weighted XGBoost	80.7817	0.1489	1.1981

The TensorFlow journey

	Tweedie Deviance	Normalized Gini	Bias Ratio
Tweedie GLM	80.6596	0.1262	1.1978
2x Weighted XGBoost	80.7817	0.1489	1.1981
Shallow Tensorflow	81.4382	0.0805	1.2157
Complex Tensforlow	80.7722	0.0969	1.2185
5x Weighted Tensorflow	80.1217	0.0717	1.2178

Final Predictions and Evaluation

- **XGBoost**
  - **%** Not able to beat GLM on important metrics – Gini is the least important metric
  - **%** Grouping was necessary to even get close which results in loss of information
  - **%** Overpredicts total risk by almost 20%
  - **+** Example risk premiums are close to GLM which is promising

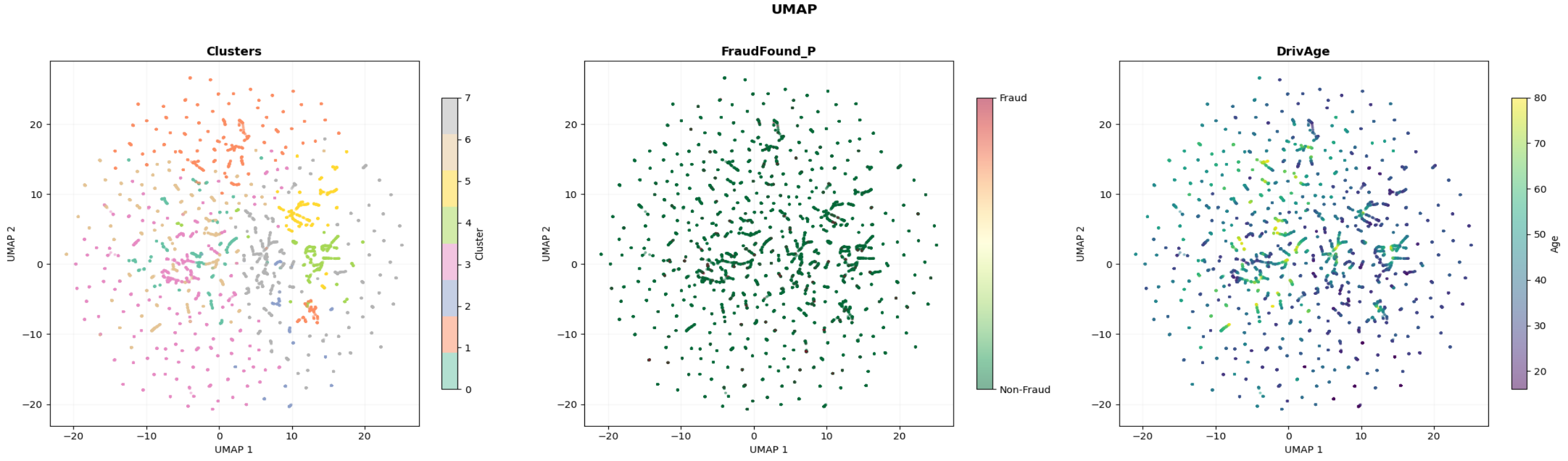
- **TensorFlow**
  - **+** Beat the GLM on Tweedie deviance which was the primary metric
  - **+** No grouping was necessary which results in more targeted risk premiums
  - **%** Overpredicts total risk by almost 22%
  - **%** Gini is significantly lower than GLM
  - **%** Example risk premiums seem quite extreme and far from baseline

	Person 1	**Person 2**	**Person 3**
Gender	Male	Male	Female
Marital Status	Other	Alone	Other
Driver Age	56	26	51
Driver Rating	51	100	147
Vehicle Age	10+	3	0
Vehicle Price	J (Medium-Low)	M (Medium)	H (Medium-Low)
GLM Risk Premium	162.67€	293.90€	227.24€
XGBoost Risk Premium	138.79€	328.65€	242.58€
TensorFlow Risk Premium	120.29€	127.32€	560.35€

- **Conclusion**
  - Due to the interpretability and transparency of the Tweedie GLM, and only a modest loss improvement using NN, the logical actuarial choice in our case is still the Tweedie GLM. However, considering the amount of data available, the project showed great promise towards the use of NN in real life pricing data.

Variational AutoEncoders

- Finding clusters and evaluating the risk
- Using variational AutoEncoders
- Finding hyper parameters using Optuna
- Clustering on the encoded data using KMEANS



Oracle — 8 Clusters

Base Fraud rate: 5.99%

Cluster	Population size	Fraud rate %	Relative to baseline	Risk group
0	1.124	**8,36%**	1,40×	**High**
1	2.556	**6,03%**	1,01×	**Normal**
2	547	**10,24%**	1,71×	**High**
3	3.575	**6,49%**	1,08×	**Normal**
4	1.089	**4,78%**	0,80×	**Normal**
5	1.028	**3,70%**	0,62×	**Low**
6	2.413	**4,72%**	0,79×	**Normal**
7	3.087	**5,93%**	0,99×	**Normal**

High risk (ratio > 1.3×) Normal risk (0.7–1.3×) Low risk (ratio < 0.7×)

Fraud risk groups summary

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Age	46	44	21	37	40	33	56	33
Sex	M	M+F	M	M	F	F	M	M
Age of Vehicle	6,64	6,30	1,51	5,61	5,66	4,89	7,72	5,24
Vehicle price*	4,90	1,58	4,96	1,29	1,58	1,35	1,32	1,46
Driver rating	76,17	130	85,28	68,70	72,64	85,37	71,69	86,22
MaritalStatus	Aline + Other**	Other	Alone	Other	Other	Alone	Other	Alone
Fraud rate	**8%**	**6%**	**10%**	**6%**	**5%**	**4%**	**5%**	**6%**
Size	1124	2556	547	3575	1089	1028	2413	3087

Vehicle price* = 0 (lowest price bracket) – 5 (highest price bracket)
Other** = Married, Divorced and widowed

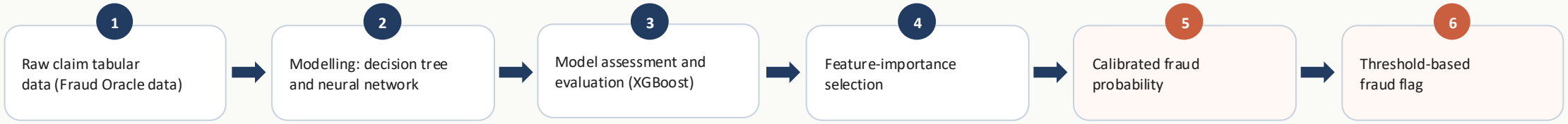
Streamlining fraud investigation

- Using claim specific tabular/image data to create machinelearning models for fraud detection

Fraud classification: objective and workflow

- **Goal:** classify fraudulent insurance claims
- **Target:** FraudFound_P
- **Model output:** predicted fraud probability
- **Use case:** prioritize claims for investigation

Workflow from claim record to investigation queue



Model evaluation and interpretability

Actual fraud rate	Mean predicted fraud probability
5.992%	5.612%

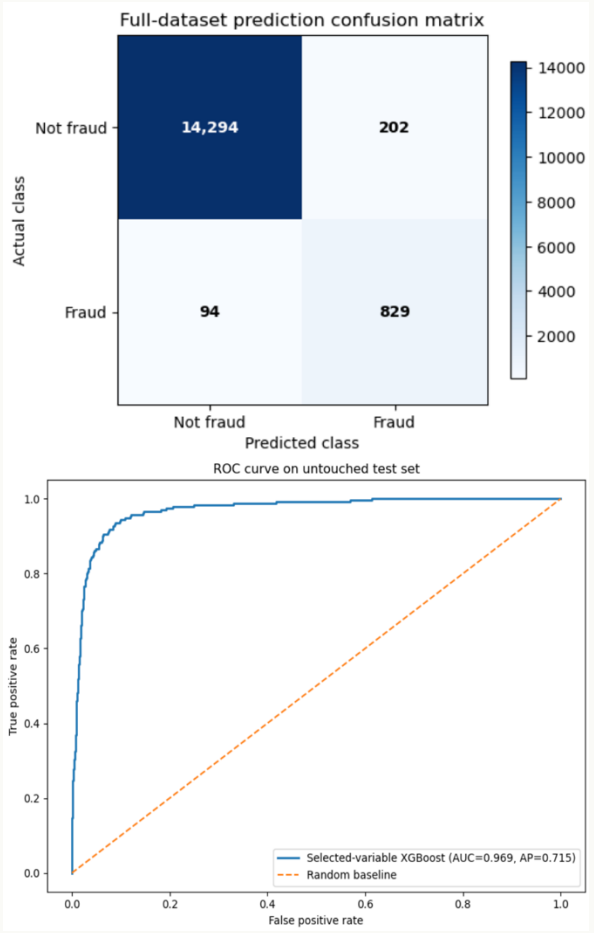
Imbalanced target: accuracy is not enough

- One hot encoding
- Feature importance threshold

Metrics

- ROC-AUC (96.91%)
- Average precision (71.5%)
- Precision (59.64%) / recall (76.88%) / F1 (67.17%)
- Confusion matrix
- Calibration

Person	Fraud probability	Predicted fraud 0/1
1	4.06%	0
2	26.26%	1
3	1.61%	0



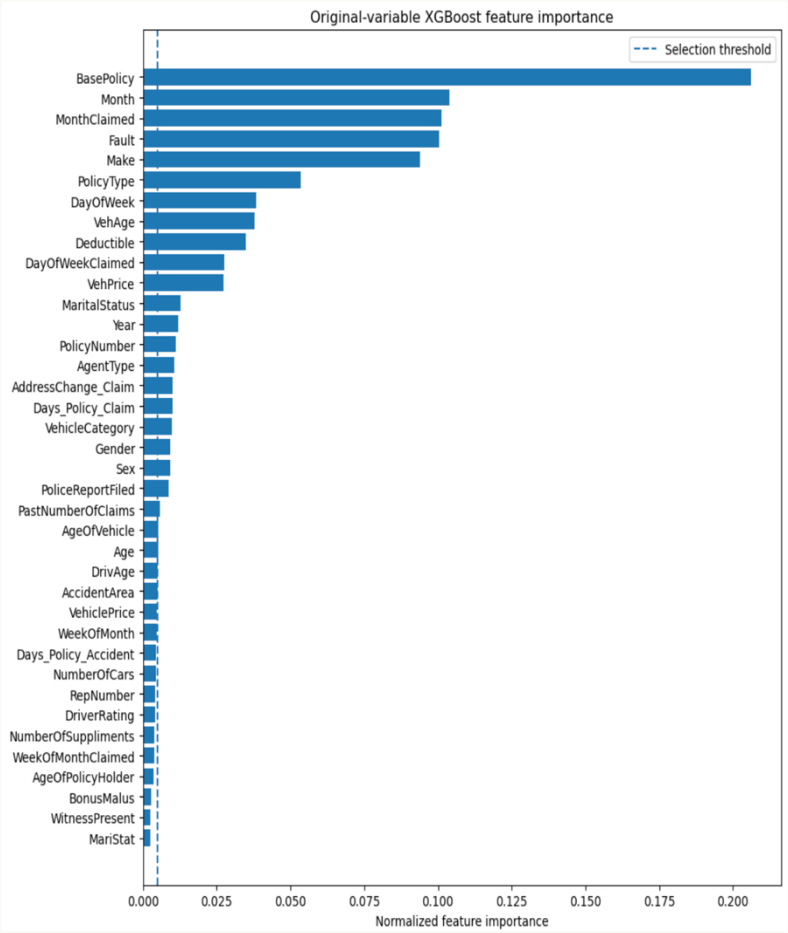


Image-based fraud classification

Can a ML model be used to classify fraud from images?

- **Data:** separate dataset of claim photos (labelled fraud / non-fraud)(fraud percent~4.6%)
- **Approach:** Anomaly detection and two stage ResNet18
- **Evaluation:** ROC-AUC, confusion matrix, F1

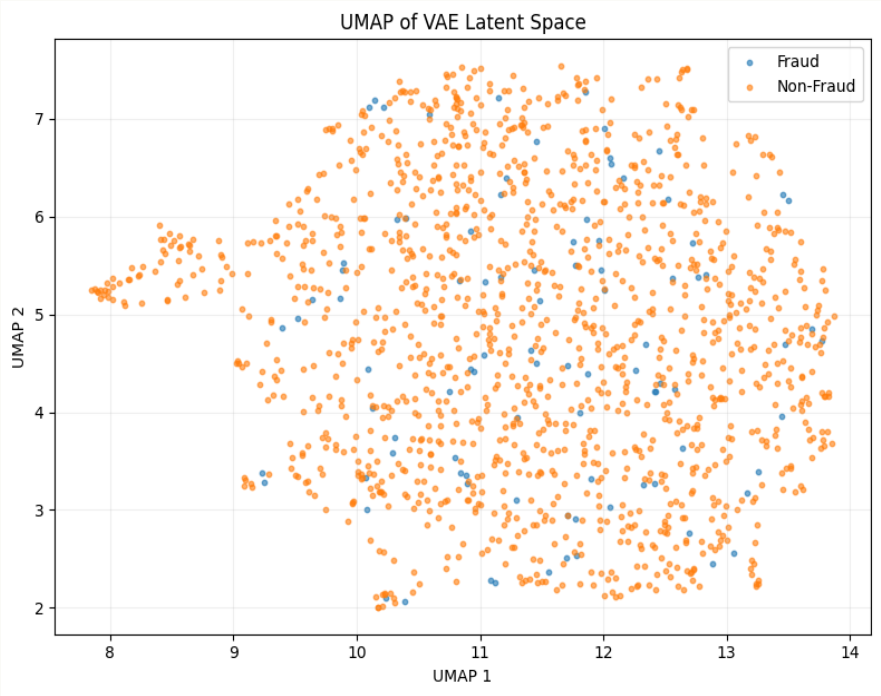
Image dataset		
Split	Fraud	Non-Fraud
Train	200	5000
Test	93	1323

Fraud

Non-Fraud

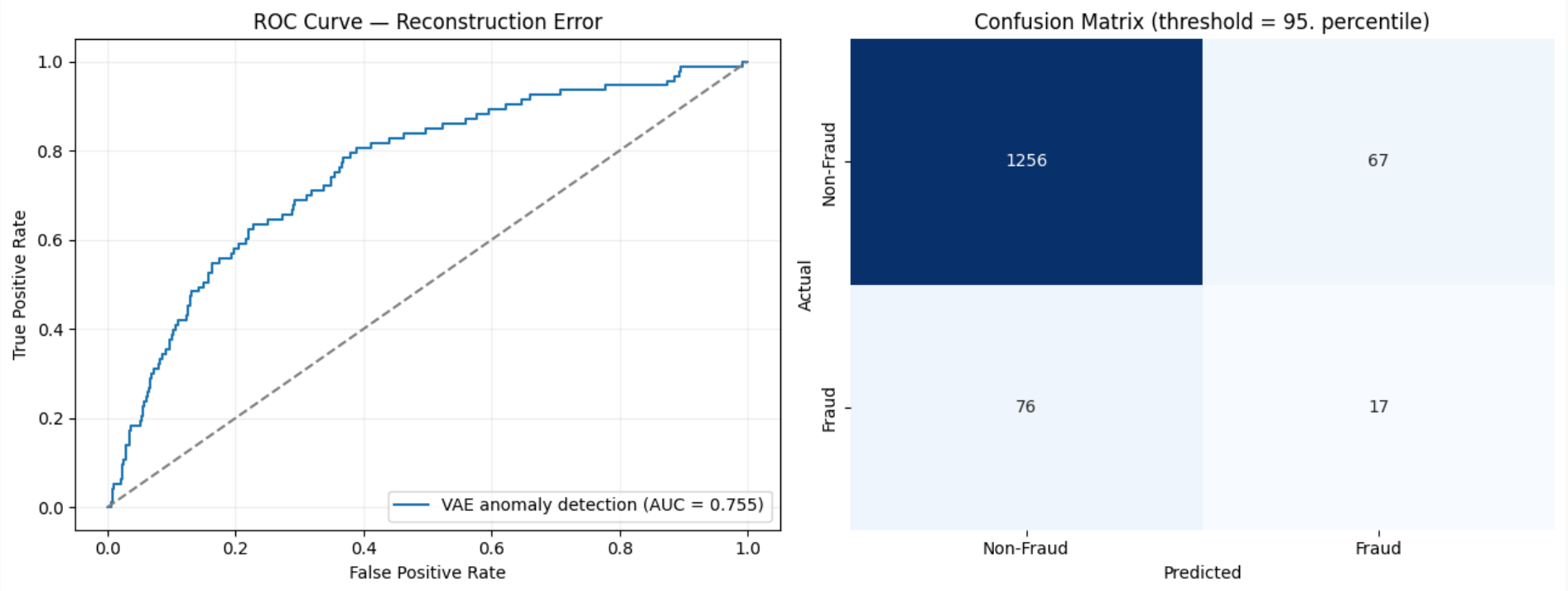
Anomaly Detection

- Convolutional VAE
- Thesis: Fraudulant claims/damages looks systematicly different
- Conclusion: A car is a car



Evaluation

- Reconstruction error is not bigger for fraud photos
- The model reconstructs fraud photos as good as non-fraud



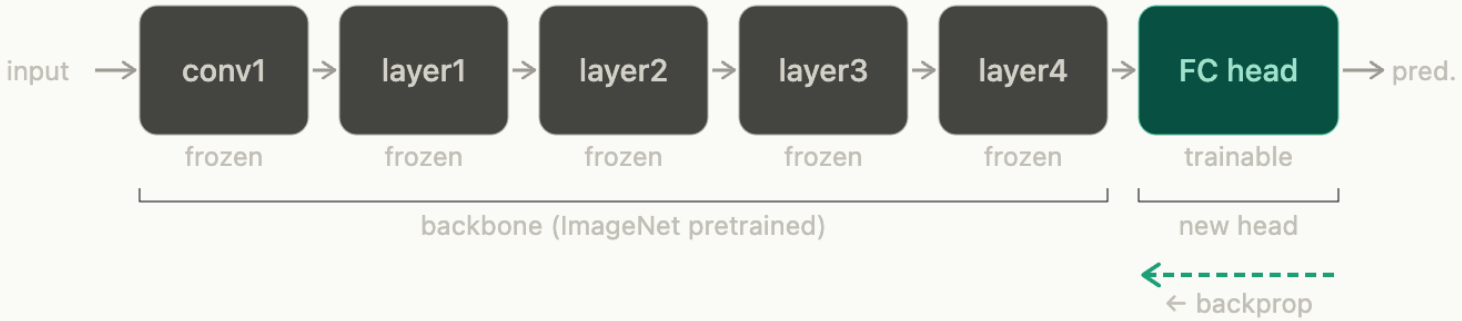
ResNet18

- Pretrained image model by Microsoft

Datasetup: Heavy data augmentation was applied during training (RandomResizedCrop, ColorJitter, RandAugment, RandomErasing) to ensure the model could generalise from only 200 Fraud training images and reduce overfitting on the minority class.

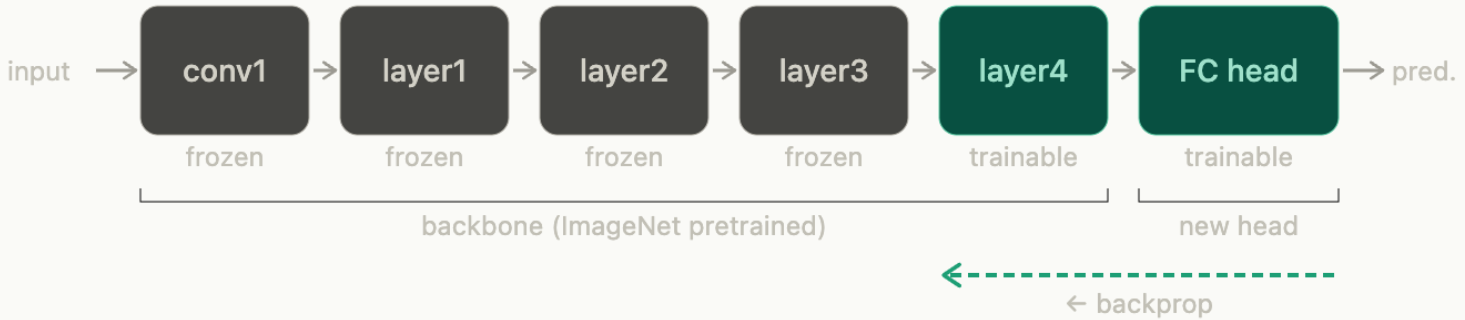
Stage 1 — Head-Only Warmup

- Backbone kept frozen so only the new randomly-initialised classification head trains first.
- Without it, large random-init gradients would immediately damage the pretrained ImageNet weights before the head has any meaningful direction.



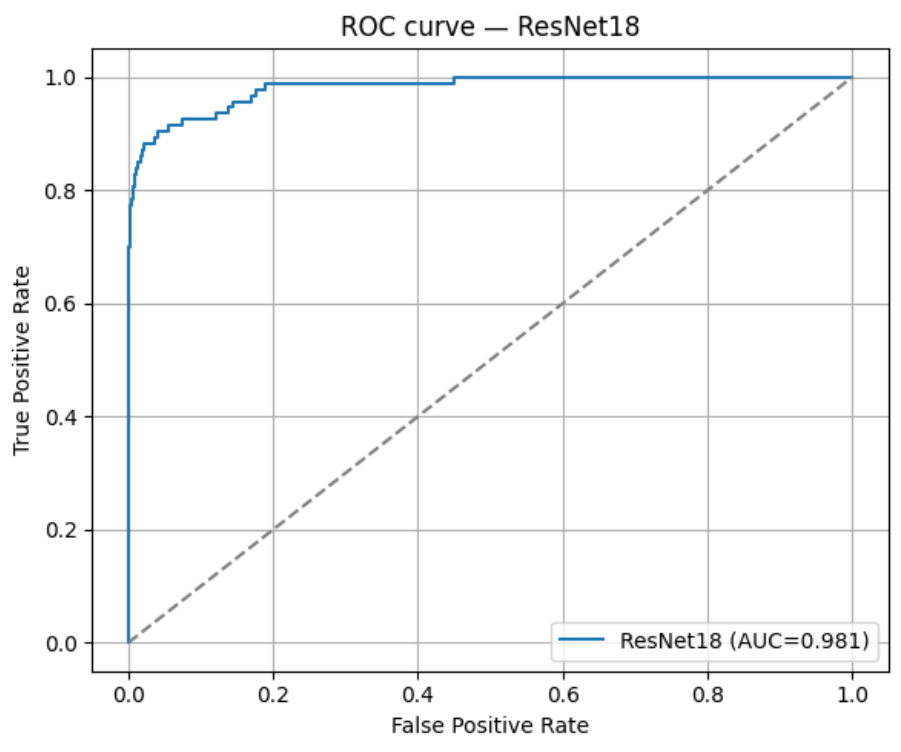
Stage 2 — Fine-Tune layer4

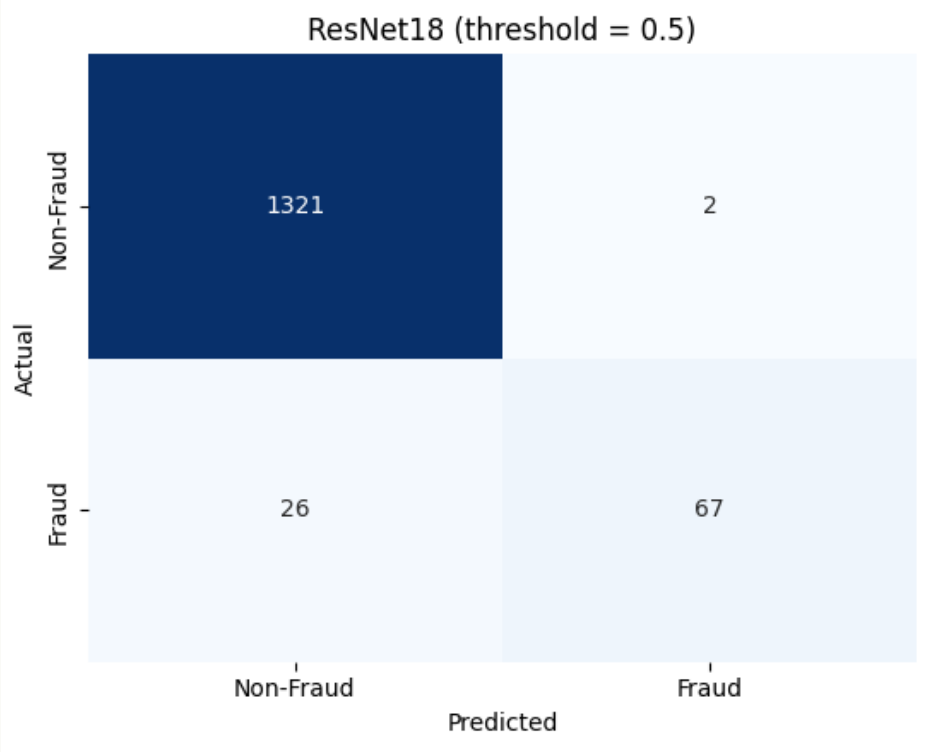
- Once the head stabilises, layer4 (the last ResNet block) is unfrozen at a 10× smaller learning rate.
- Most of the quality gain comes from here, high-level features shift from generic object recognition toward the specific patterns in damaged-car photos.



ResNet18

- Pretrained image model by Microsoft





Final fraud detection evaluation

Tabular based model

- $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 0.848$

Image based model

- $F1 = 0.827$

Conclussion: In general our models are able to help claims handlers to focus investigative resources where they matter most. Results are promising, but performance could be further improved with larger and more diverse training data, additional image quality controls, and integration between the two models.