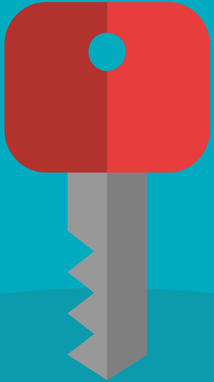
Vehicle Insurance Fraud Detection

Presented by Group 19:
Bian Tong A0234470L
Kang Xuan Hui Victoria A0241262R
Ke Jiayi A0241155N

01

Motivation




$29 Billion
Annual Premium Leakage in the
United States alone

Types of Vehicle Fraud

& Why is it Challenging?

Types

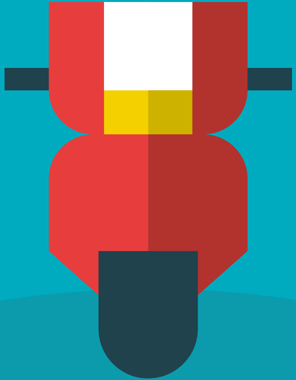
1. Staged Accidents
2. Exaggeration of Damages or Injuries
3. Vehicle Theft Fraud
4. Ghost Claims
5. Identity Fraud

Why is it challenging?

1. Complexity of Fraudulent behaviour
2. Data Overload
3. Lack of clear fraud indicators
4. Human element and Subjectivity
5. Legal and Ethical Concerns

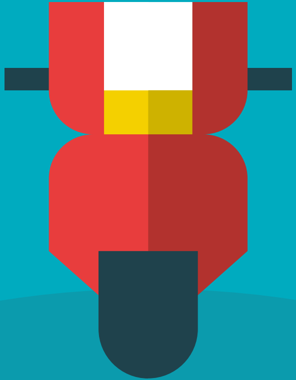
02

Problem Statement



02

Problem Statement



Given the complexity of the vehicle insurance claims, how can we detect fraudulent claims in a sea of legitimate ones?

03

Our Dataset

Vehicle Insurance Claims in the United States circa 1990s

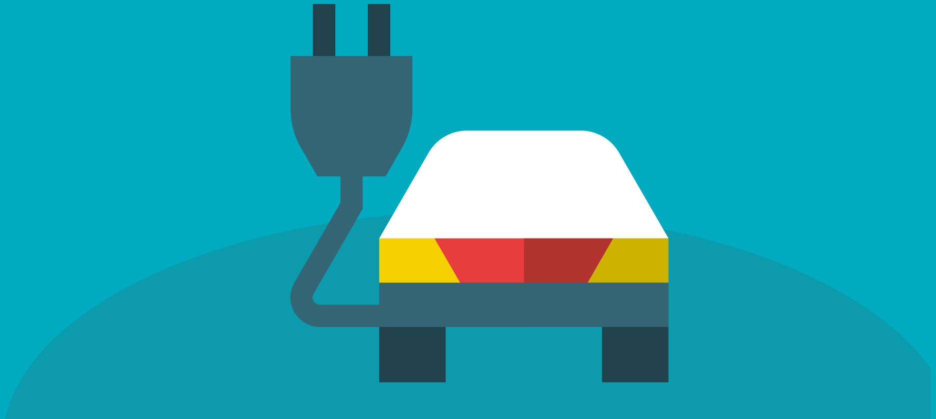
Fraud_oracle.csv

- From Vehicle Insurance Industry
- 9.41 Kaggle usability

- 33 features
- 15,420 records
- 6% (923) fraudulent.

04

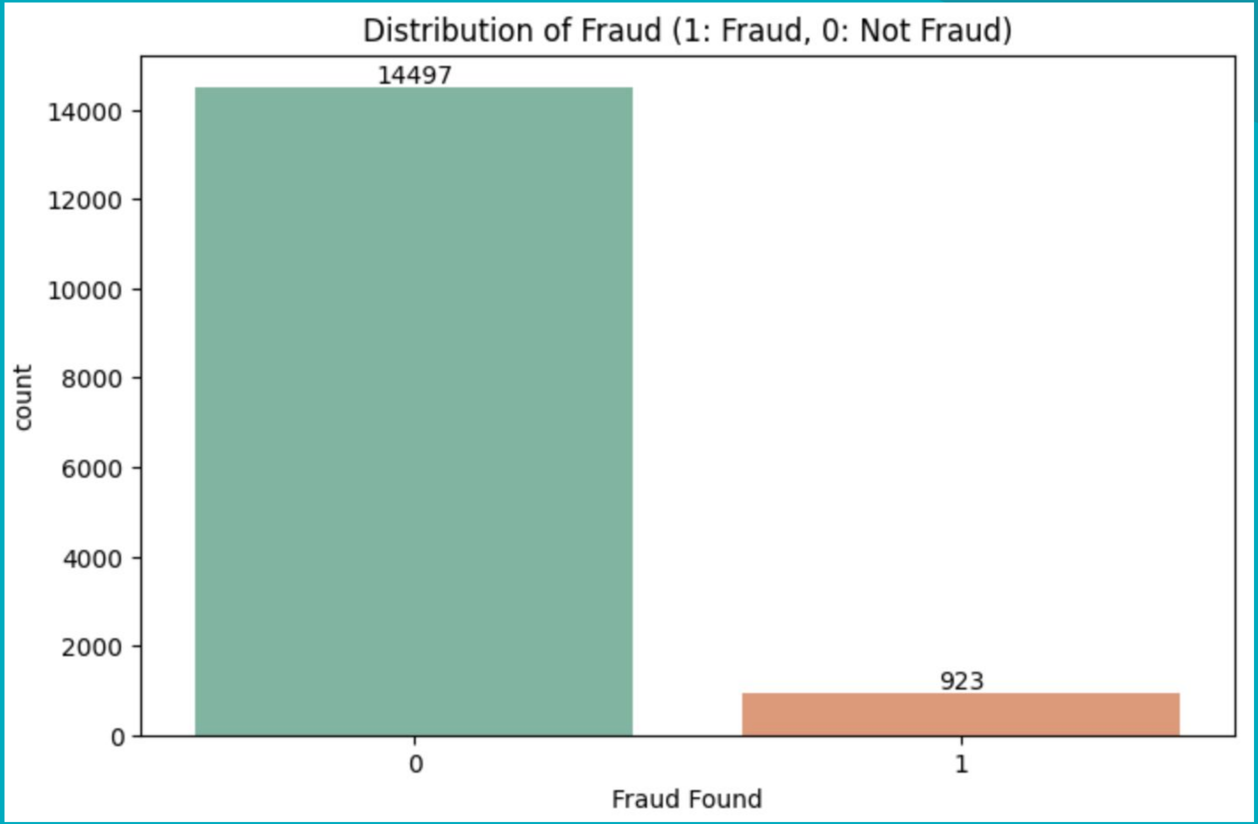
Data Preprocessing & Modelling



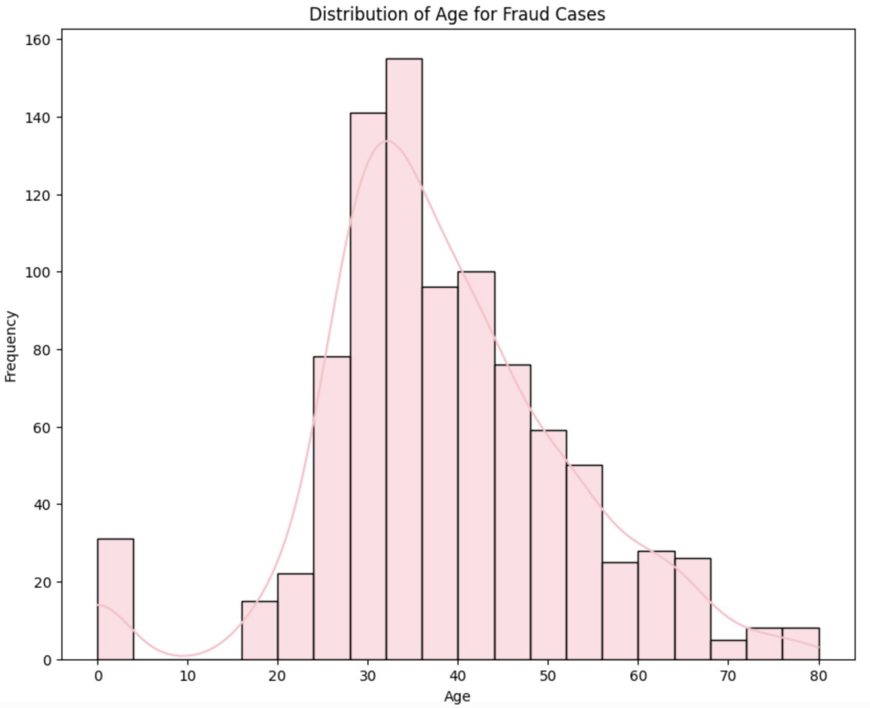
Data Preprocessing - Convert Data Type

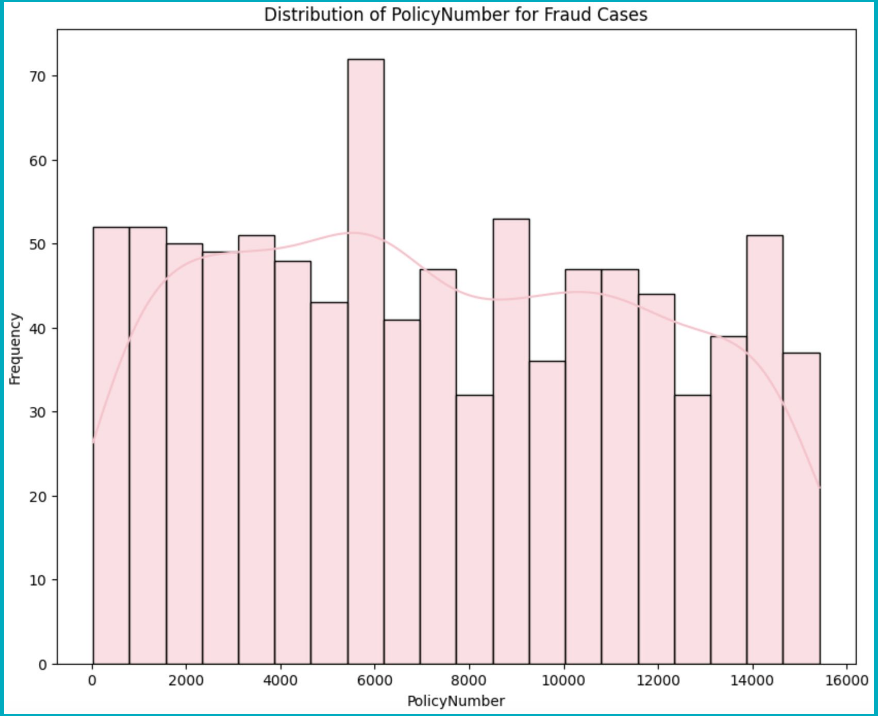
	DataType	UniqueValues
WitnessPresent	object	2
AgentType	object	2
FraudFound_P	int64	2
AccidentArea	object	2
PoliceReportFiled	object	2
Fault	object	2
Sex	object	2
Year	int64	3
BasePolicy	object	3
VehicleCategory	object	3
PastNumberOfClaims	object	4
Days_Policy_Claim	object	4
DriverRating	int64	4
Deductible	int64	4
MaritalStatus	object	4
NumberOfSuppliments	object	4
NumberOfCars	object	5
WeekOfMonthClaimed	int64	5
Days_Policy_Accident	object	5
AddressChange_Claim	object	5
WeekOfMonth	int64	5
VehiclePrice	object	6
DayOfWeek	object	7
DayOfWeekClaimed	object	8
AgeOfVehicle	object	8
PolicyType	object	9
AgeOfPolicyHolder	object	9
Month	object	12
MonthClaimed	object	13
RepNumber	int64	16
Make	object	19
Age	int64	66
PolicyNumber	int64	15420

Data Preprocessing - SMOTE

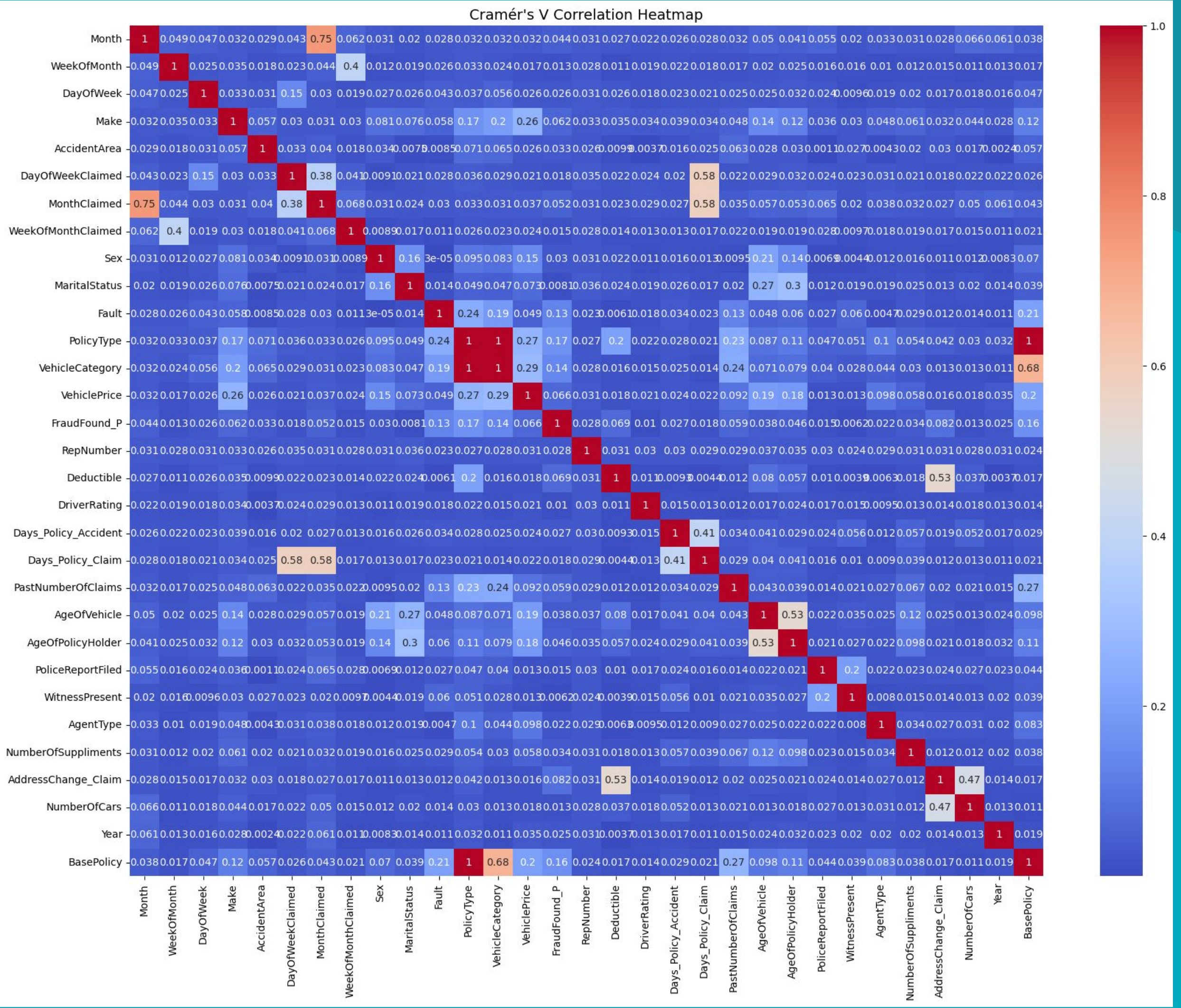


Data Preprocessing - Drop Useless Columns





Data Preprocessing - Drop Highly Correlated Columns



Data Preprocessing - Feature Encoding

Binary

Label Encoding to
0 and 1 indicating
True and False

Ordinal

Specified mappings to
retain the ordering.

```
month_mapping = {'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4, 'May': 5, 'Jun': 6,
                 'Jul': 7, 'Aug': 8, 'Sep': 9, 'Oct': 10, 'Nov': 11, 'Dec': 12}

day_of_week_mapping = {'Monday': 1, 'Tuesday': 2, 'Wednesday': 3,
                       'Thursday': 4, 'Friday': 5, 'Saturday': 6, 'Sunday': 7}
```

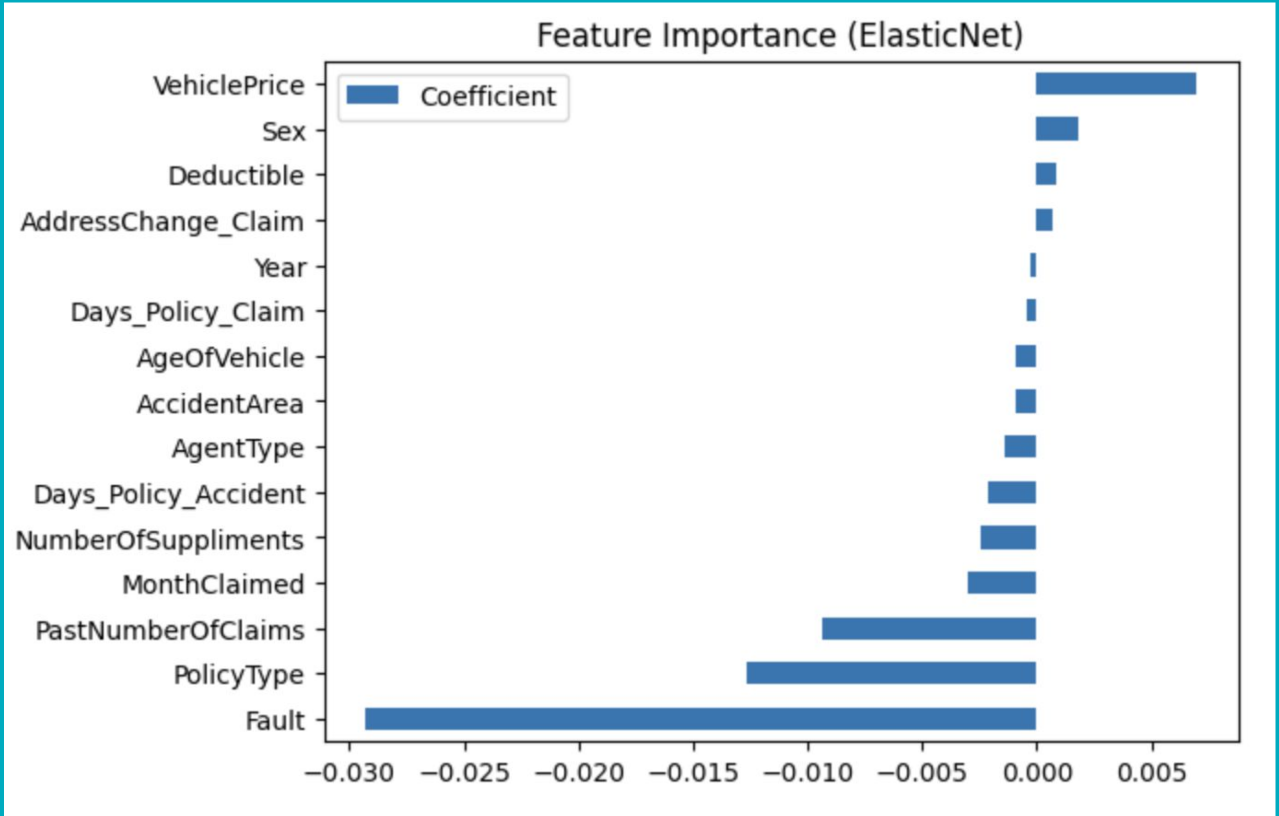
Nominal

Auto Label Encoding

Data Preprocessing - Others

Scaling	Train-Val-Test Split	Drop NA
Standard Scaler	56%-14%-30%	Make sure no NA before training

Feature Selection - ElasticNet



Modelling - SMOTE vs Non-SMOTE

Logistic Regression

Baseline model, limited in capturing non-linear relationship.

Random Forest

Ensemble model, good at handling non-linear relationship.

XGBoost

Highly efficient in handling imbalanced dataset.

CatBoost

Handling of categorical variables, robust against overfitting.

Neural Network

Capture complex patterns, sensitive to data imbalance.

KNN

Simple non-parametric, less effective with imbalance data.

Non-SMOTE

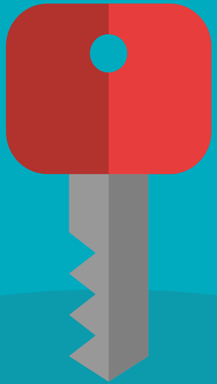
Models	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression (Baseline)	74.15%	11.35%	48.84%	73.92%
KNN	93.98%	47.62%	7.75%	64.02%
Random Forest	93.28%	28.95%	8.53%	79.15%
XGBoost	93.38%	20.83%	3.88%	6.54%
CatBoost	82.91%	19.70%	60.47%	83.99%
Neural Network	94.03%	0.00%	0.00%	80.60%

SMOTE

Models	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression (Baseline)	67.67%	11.61%	66.67%	72.58%
KNN	80.92%	14.18%	43.41%	69.54%
Random Forest	90.23%	25%	31.78%	79.45%
XGBoost	89.02%	22.73%	34.88%	82.61%
CatBoost	88.74%	21.78%	34.11%	82.86%
Neural Network	80.13%	15.91%	54.26%	82.47%

05

Results



Model Results After Hyperparameter Tuning

Models	ACCURACY	PRECISION	RECALL	ROC-AUC
LOGISTIC REGRESSION	66.33%	10.97%	65.12%	72.67%
KNN	79.16%	15.03%	53.49%	75.04%
RANDOM FOREST	89.12%	22.40%	33.33%	81.74%
XGBOOST	88.78%	21.19%	32.13%	82.59%
CATBOOST	70.73%	13.71%	73.64%	80.60%
NEURAL NETWORK	-	-	-	-

Integration of the model into the business process

CLAIM PRE-SCREENING

Automatically analyse incoming insurance claims and assign a fraud risk score to each submission.

ENHANCED FRAUD DETECTION WORKFLOW

Supplement traditional rule-based approaches with advanced, data-driven insights.

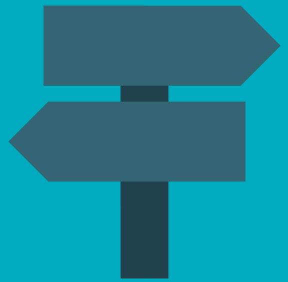
REAL-TIME DECISION SUPPORT

Provide instant fraud risk assessments to identify potentially fraudulent claims early in the process.

Additional Insights from Experiments

Reducing Decision Thresholds

Trade-offs with precision, accuracy and ROC-AUC



Limitations

OUTDATED DATASET

Dataset used is nearly 30 years old. Advancements in technology have transformed fraud detection and monitoring methods

GEOGRAPHICAL BIAS

Limited generalisability of the findings to other regions

HUMAN OVERSIGHT

Fraudulent claims often involve complex and nuanced contexts that automated systems may not fully capture.

Conclusion

CATBOOST

Can identify fraudulent insurance claims with a high recall score.

RELEVANCE

Despite advancements, claims data have remained consistent, ensuring the applicability of our findings to modern fraud detection scenarios.

FUTURE WORKS

Focus on addressing challenges in detecting minority class instances more effectively

Thank You