Unmasking Deception

Predictive Analytics in Automobile Insurance Fraud

By: Hazem Medhat

Table OF Content

Introduction

Problem Statment

- Insurance fraud is a significant issue, costing the industry billions annually.
- Identifying fraudulent claims is challenging due to complex patterns and diverse data.
- Need for a robust system to detect fraud efficiently, reducing false positives and improving claim processing.

Business Overview

- The insurance company processes thousands of claims annually, covering various vehicle types and policy categories.
- Key stakeholders: policyholders, claims adjusters, and fraud investigators.

Project Objective

- Develop a predictive model to identify potentially fraudulent car insurance claims.
- Enhance decision-making for claims processing using data-driven insights.
- Minimize financial losses due to fraud while maintaining efficient claim handling

Data

Data Set

- The dataset is an automobile insurance dataset “carclaims”, which is publically available and is provided by Angoss Knowledge Seeker.
- Size: 15,420 records, 33 attribute (32 Feature , 1 Target).
- Kaggle Link : CarClaims

Meta Data

Column Name	Type	Description
Month	Categorical	Month of the accident
WeekOfMonth	Ordinal	Week of the month when the accident occurred
DayOfWeek	Categorical	Day of the week of the accident
Make	Categorical	Vehicle manufacturer
AccidentArea	Binary	Location of the accident
DayOfWeekClaimed	Categorical	Day of the week the claim was filed
MonthClaimed	Categorical	Month the claim was filed

Meta Data

Column Name	Type	Description
FraudFound	**Binary**	**Target variable indicating fraud (Yes) or legitimate (No) claim**
WeekOfMonthClaimed	Ordinal	Week of the month when the claim was filed
Sex	Binary	Gender of the policyholder
MaritalStatus	Categorical	Marital status of the policyholder
Age	Numeric	Age of the policyholder
Fault	Binary	Who was at fault for the accident
PolicyType	Categorical	Type of insurance policy
VehicleCategory	Categorical	Category of the vehicle

Meta Data

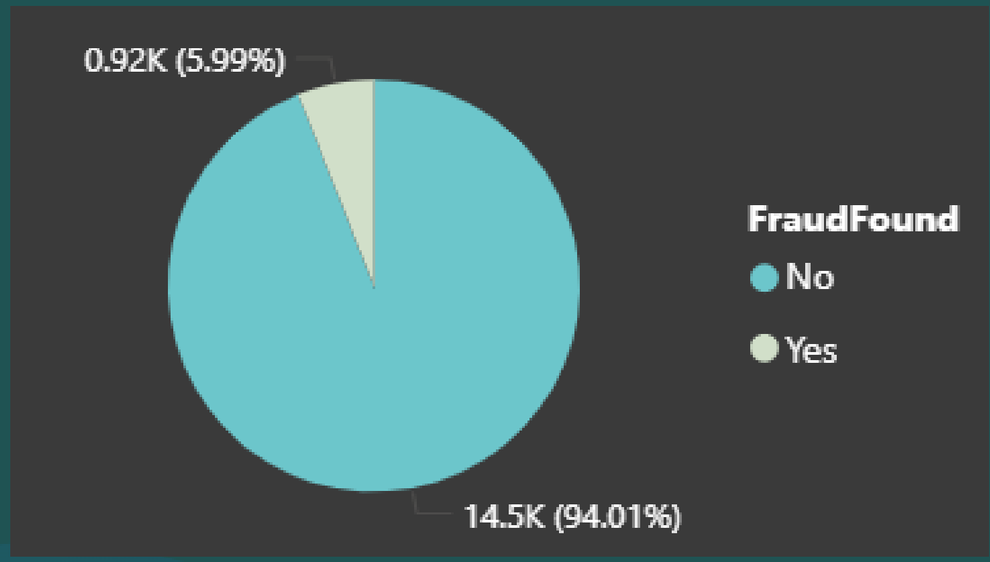
Column Name	Type	Description
PolicyNumber	Numeric	Unique identifier for the policy
RepNumber	Numeric	Representative number handling the claim
Deductible	Numeric	Deductible amount for the claim
DriverRating	Ordinal	Rating of the driver
Days:Policy-Accident	Ordinal	Days between policy start and accident
Days:Policy-Claim	Ordinal	Days between policy start and claim filing
PastNumberOfClaims	Ordinal	Number of past claims

Meta Data

Column Name	Type	Description
AgeOfPolicyHolder	Ordinal	Age range of the policyholder
PoliceReportFiled	Binary	Whether a police report was filed
WitnessPresent	Binary	Whether a witness was present
AgentType	Binary	Type of agent handling the claim
NumberOfSuppliments	Ordinal	Number of supplements filed
AddressChange-Claim	Ordinal	Time since address change before claim
NumberOfCars	Ordinal	Number of cars insured
Year	Numeric	Year of the claim
BasePolicy	Categorical	Base policy type

Data Challenges

- Imbalanced Data: Fraudulent claims are rare (e.g., 6% of claims are fraudulent).



- Missing Values: Age values of 0 for some records, indicating potential data entry errors.

Analysis & Insights

Data Cleaning & Transformation

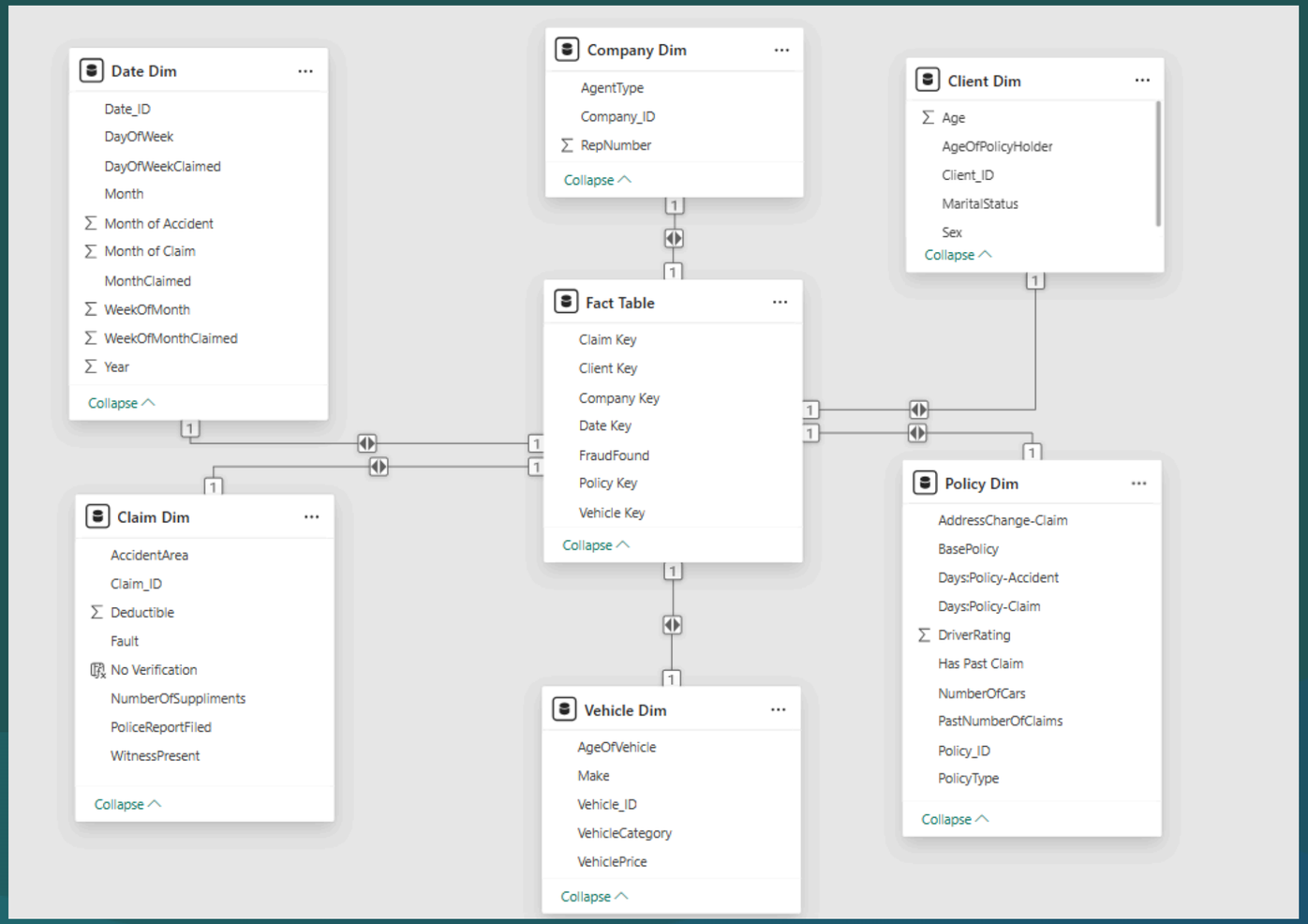
- Handling Missing Values: Replaced Age = 0 with median age .

- Transformation :
  - “Has Past Claims” → New derived binary column to indicate if the client has provided a previous claim or not.

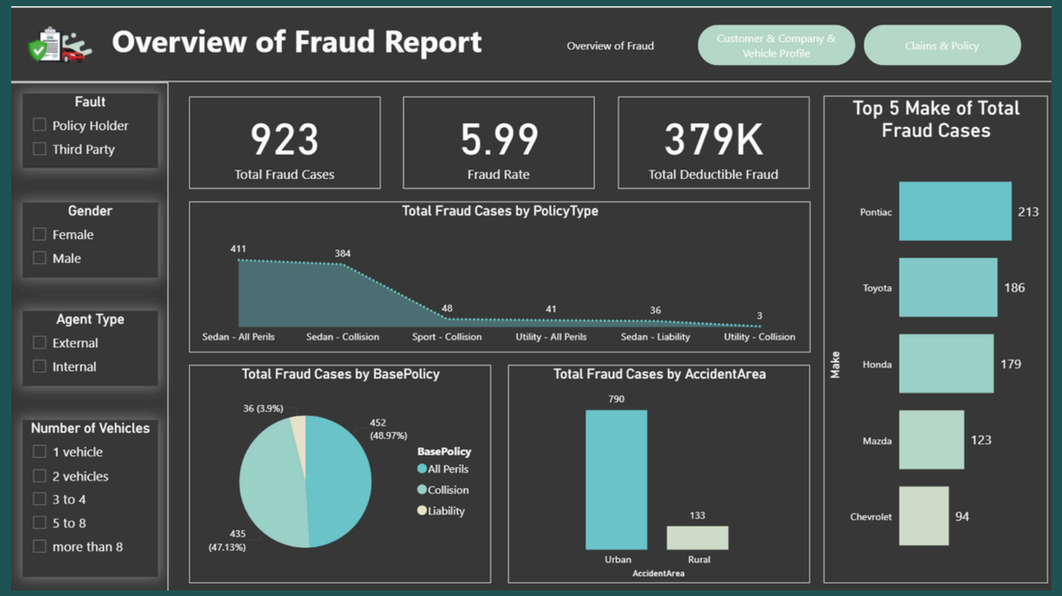
  - “No Verification” → New derived binary column to indicate if the client has provided any verification to police of the accident or have a witnesses of it

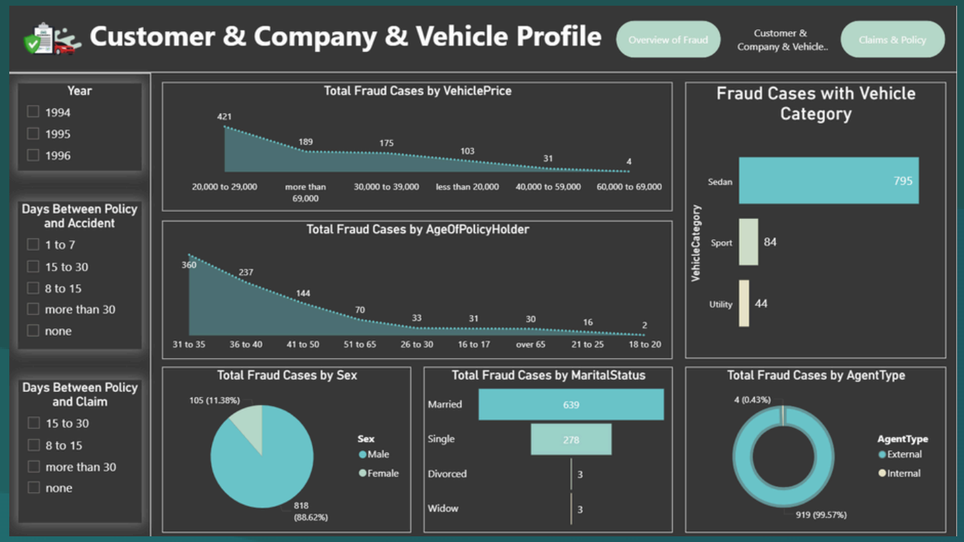
Data Modeling

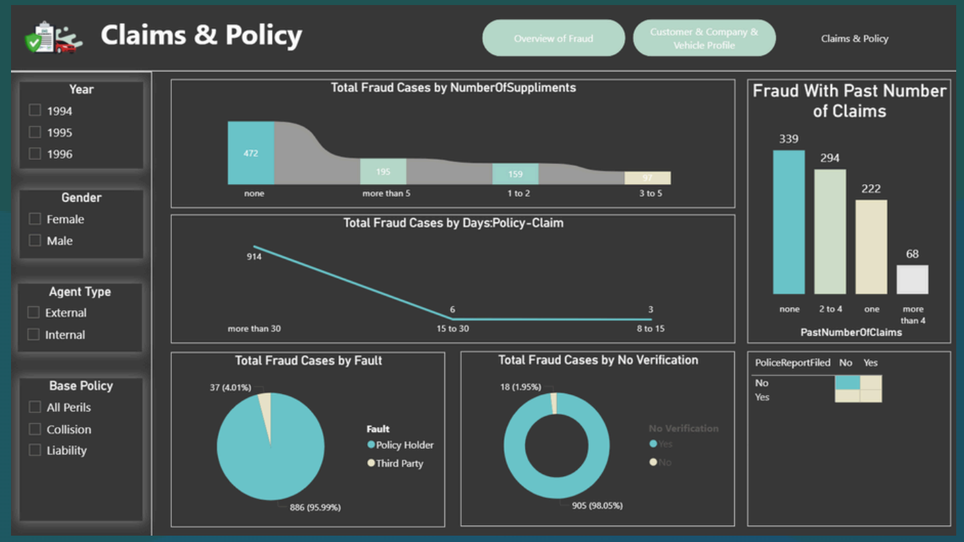
- Transform data from one table to Star Schema



Visuals & Insights







Visuals & Insights

- Fraudulent claims are most prevalent under the "Sedan - All Perils" policy type, indicating a need for enhanced scrutiny in this category.
- Urban areas exhibit a significantly higher incidence of fraudulent claims compared to rural areas, suggesting a focus on urban claim patterns.
- The majority of fraudulent claims are associated with policyholders aged 31 to 35, pointing to this age group as a key area for fraud prevention.
- Married policyholders account for the largest share of fraudulent claims, suggesting targeted monitoring of this demographic.
- Claims filed within 8 to 15 days of the policy are the most common for fraud, indicating rapid filings as a potential red flag.
- Fraudulent claims with no supplements are the most frequent, highlighting the importance of enforcing stricter documentation requirements.

Prediction Models & Results

1-Feature Engineering

Add 2 New Columns :

- Month diff → Difference between MonthClaimed and Month

- No Verification → Derived feature that shows if the claim has any verification (1 if PoliceReportFiled = No and WitnessPresent = No, else 0).

2-Encoding

Apply 2 Types of encoding :

- Target → for categorical columns with many categories .

- Label → for categorical columns with few categories.

3- Scaling

- Apply Standard Scaling

4- Handle Class Imbalance

- The dataset has a 94:6 imbalance (FraudFound: 14,497 No vs. 923 Yes), which can bias models toward the majority class.

- SMOTE (Synthetic Minority Oversampling Technique) generates synthetic samples for the minority class (fraudulent claims) by interpolating between existing samples, preserving feature relationships. This helps models learn fraud patterns without overfitting to legitimate claims.

5- Models and Results

Model	Accuracy Result
SVM	83%
Logistic Regression	67%
XGBosst	**94%**
Naive Bayes	65%
Decision Tree	88.7%
Random Forest	93.7%

Conclusion

Conclusion

- Developed a robust model to detect fraudulent insurance claims with high accuracy.
- Identified key risk factors: young drivers, urban areas, and claim delays.

Future Work

- Model Deployment and Incorporate real-time data and advanced anomaly detection techniques.

Thank **You!**