Modeling Car Insurance Claim Outcomes Using Machine Learning

1. Background: Insurance Companies and Risk Assessment

Insurance companies operate in a highly competitive and risk-sensitive environment. One of their main challenges is accurately pricing insurance policies and estimating the probability that a customer will make a claim. This process requires analyzing large amounts of customer data related to personal characteristics, driving behavior, and vehicle information.

In many countries, car insurance is a legal requirement for driving on public roads, which makes the automobile insurance market extremely large and financially significant. As a result, even small improvements in claim prediction accuracy can lead to substantial cost savings and better risk management decisions.

According to industry research, insurance companies invest heavily in data analytics and machine learning techniques to improve underwriting and pricing strategies
(Source: Accenture – Machine Learning in Insurance).

2. Project Context

In this project, we assume the role of a data analyst working with On the Road Car Insurance, a fictional car insurance company.
The company aims to improve its decision-making process by predicting whether a customer is likely to file an insurance claim during the policy period.

However, the company has limited expertise and infrastructure for deploying complex machine learning models. Therefore, they requested a simple yet effective solution, focusing on identifying the most informative features and building a model that is easy to interpret and deploy.

To achieve this, the company provided a real customer dataset in CSV format named:

car_insurance.csv

3. The Dataset

The dataset contains customer-level information collected by the insurance company. It includes demographic attributes, driving history, and vehicle-related variables.

Main Characteristics of the Data

- 10,000 records
- 16 features describing each customer
- A binary target variable indicating whether a claim was made

Dataset Columns

COLUMN	DESCRIPTION
id	Unique client identifier
age	Client's age group
gender	Client's gender
driving_experience	Years of driving experience
education	Level of education
income	Income category
credit_score	Credit score (between zero and one)
vehicle_ownership	Whether the client owns the vehicle
vehicle_year	Vehicle registration year
married	Marital status
children	Number of children
postal_code	Postal code
annual_mileage	Miles driven per year
vehicle_type	Type of vehicle
speeding_violations	Number of speeding violations
duis	Number of DUI incidents

past_accidents	Number of past accidents
outcome	Whether a claim was made

4. Project Objective

The main goal of this project is to:

Predict whether a customer will make an insurance claim or not.

Achieving this helps insurance companies:

- Select lower-risk clients
- Improve underwriting decisions
- Reduce expected losses
- Build more reliable pricing strategies

5. Methodology and Modeling

The project was divided into three main stages:

1. Data Analysis and Preparation

- Encoding categorical variables into numerical form
- Handling missing values
- Selecting the most relevant features for modeling

2. Machine Learning Modeling

Several machine learning models were tested, including:

- Logistic Regression
- Decision Tree
- K-Nearest Neighbors
- Random Forest and ensemble methods

After evaluation, the Decision Tree classifier achieved the best balance between accuracy and interpretability, with an accuracy of approximately 85%.

The final model was trained using:

- Entropy as the splitting criterion
- Maximum tree depth of five
- Minimum samples per split equal to five

This ensured a robust yet simple model suitable for production use.

6. Application Development

To make the model usable in a real-world scenario, an interactive web application was developed using Streamlit.

The application allows users to:

- Enter customer information through a simple interface
- Instantly receive a prediction about claim likelihood

Example Scenario

- A young client with limited driving experience and low income is predicted to be likely to make a claim
- A more experienced driver with stable income and clean driving history is predicted to be unlikely to make a claim

This enables insurance companies to quickly evaluate applications and make informed decisions.

7. Conclusion

In this project, we successfully:

- Analyzed real-world car insurance data
- Built a reliable machine learning model for claim prediction
- Identified key risk-related features
- Developed a practical web application for decision support

The final result is a fully functional prediction system that can help insurance companies assess customer risk with high accuracy and minimal complexity.

This project demonstrates how data analysis and machine learning can be effectively applied to solve real business problems in the insurance industry.