Executive Summary – Predicting Car Insurance Claims

Team Name: Data Dynamos
Date: June 23, 2025.

Overview

This analysis aims to develop predictive models to determine whether a car insurance policyholder will file a claim within the next 6 months. Using the provided dataset, we aim to uncover key factors influencing claim behaviours and train models to accurately predict future claims.

Approach

The dataset underwent extensive pre-processing to ensure data quality and model readiness. This included checking for missing values, and addressing class imbalance.

We began with an Exploratory Data Analysis (EDA) to better understand the structure of the data:

- Univariate analysis was used to assess the distribution of individual features.
- Bivariate analysis examined the relationship between each feature and the target variable is_claim.

After EDA, we dropped irrelevant columns, applied one-hot encoding to categorical features, and scaled numerical variables for consistency.

To reduce dimensionality and minimize high correlations among features, we applied Principal Component Analysis (PCA) based on the correlation matrix. The first 15 principal components were retained, capturing approximately 79.62% of the total variance, ensuring that most of the important information in the dataset was preserved for modelling.

The dataset was split into training and testing sets using a 70/30 ratio. To address the significant class imbalance, we applied SMOTE, a resampling technique designed to generate synthetic minority class samples, thereby improving the model's ability to detect rare claim cases.

Based on our insights, three classification models were tested:

1. Logistic Regression – for its simplicity and interpretability.
2. Decision Tree – for capturing non-linear patterns in the data.
3. Random Forest – for its ensemble power and ability to improve prediction accuracy while controlling overfitting.

All models were evaluated using a combination of accuracy, precision, recall, F1-score, and AUC to ensure a balanced assessment of performance.

Findings

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.8067	0.0866	0.2166	0.1209	0.5743
Logistic Regression	0.5850	0.0775	0.5117	0.1341	0.5697
Decision Tree	0.5726	0.0704	0.4738	0.1222	0.5317

A summary of the results is presented below:

- **Random Forest** achieved the highest accuracy (80.67%) and precision (8.66%), but had low recall (21.66%). This indicates that while it performed well overall, it still struggled to identify a large proportion of actual claimants.
- **Logistic Regression** had the highest recall (51.17%) among all models, which is valuable in identifying more actual claimants. However, its precision (7.75%) and accuracy (58.50%) were relatively low.
- **Decision Tree** performed slightly below logistic regression, with recall of 47.38% and F1 score of 12.22%, showing that it detected some claims but also produced many false positives.

Overall, Random Forest achieved the highest accuracy and precision, Logistic Regression outperformed in recall and F1 score, while the Decision Tree model delivered moderate results and retained a high level of interpretability.

However, the relatively low AUC values across all models (0.5317 to 0.5743) show that the models have difficulty distinguishing between claimants and non-claimants. This could be due to the class imbalance and/or possible information loss from PCA. To improve performance, further feature engineering, model tuning, or trying other algorithms may be needed in the future.

Interpretation

1. For Logistic Regression; PC12 had the strongest negative effect on claim probability (–0.171, $p < 0.001$), PC13 showed a strong positive effect (+0.127, $p < 0.001$), PC11 and PC14 also had significant negative effects while PC3, PC5, and PC15 were not statistically significant.
2. Decision tree: The decision tree relied primarily on two principal components (PC11 and PC12) to split the data, with the majority of claim predictions concentrated in a single branch, reflecting the class imbalance in the dataset.
3. The Random Forest model with 500 trees relied most on three key components (PC13, PC12, and PC11) to make its predictions. These components captured the most useful patterns in the data, helping the model distinguish between those likely and unlikely to file a claim even though we can't interpret them directly.

Recommendation

Based on our evaluation, we recommend deploying the Logistic Regression model for initial implementation. Although it had lower overall accuracy, it achieved the highest recall, which is critical in identifying policyholders who are likely to file a claim.

Also, Logistic Regression offers high interpretability, making it easier to explain to business stakeholders and justify underwriting decisions. This transparency is especially important in regulated industries like insurance.

Business Insights

Accurately predicting which policyholders are likely to file a claim in the next six months enables insurers to make better decisions around risk-based pricing, underwriting, and resource allocation. High-recall models like Logistic Regression can help flag potential high-risk customers for further review or tailored premium adjustments. While precision remains low across all models, flagging more true positives is preferable in this context, as missing a claimant could result in unexpected losses. These insights can also support fraud detection, targeted communication, and customer segmentation strategies.