Car Insurance Claim Prediction

Project 2

José Cunha – up201905451
Mariana Teixeira – up201905705
Raquel Carneiro - up202005330

Specification of the Problem

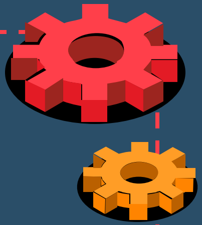
The Dataset contains information on policyholders having the attributes like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the car, power, engine type, etc, and the target variable indicating whether the policyholder files a claim in the next 6 months or not.

Goal:

train an AI to **predict** whether the policyholder will file a claim in the next 6 months or not, based on information on policyholders.

Related Work:

- https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification?select=train.csv



Algorithms

- Neural networks

  Machine learning algorithm; Learning patterns and making predictions with remarkable accuracy.

- Decision Trees

  Intuitive models that recursively split data based on features, enabling clear rule-based decision-making and effective pattern recognition.

- Gradient Boosting

  An ensemble learning technique that combines weak models sequentially, learning from their mistakes to create a strong predictive model capable of handling complex relationships and achieving high accuracy.

Other algorithms and tools

- Logictic Regression

  Commonly used to predict insurance based on features, forecasting binary outcomes and probabilities.

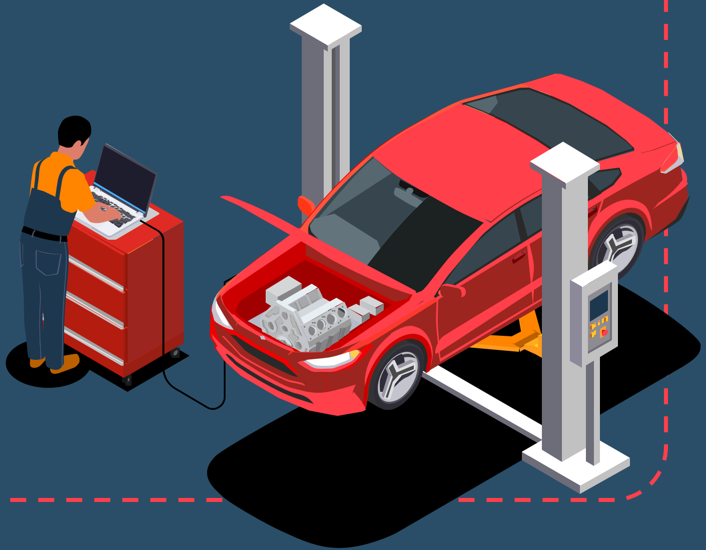
- Random Forests

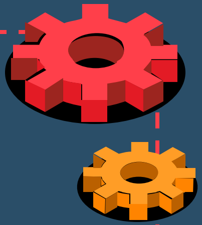
  Ensemble learning algorithm that combines multiple decision trees to improve accuracy and robustness.

- XGBoosting (eXtreme Gradient Boosting)

  Combines multiple weak models to create a strong model.

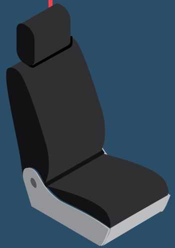
- Libraries:
  - Jupiter Notebook
  - Pandas library
  - Mathplotlib library
  - scikit-learn

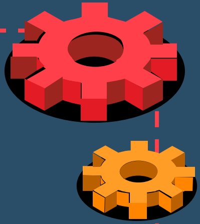




Data pre-processing

- Cleaning and transforming original dataset
  - Removing missing values, duplicates
  - Ensure there were no values out of ranges or anormal – normalize data
  - Handling outliers
  - Get samples – under sampling and over sampling
  - Characterize variables, describe data
  - Encoding categorial variables??
- Univariate and bivariate analysis
- Feature scaling.





Developed Models

Logictic Regression

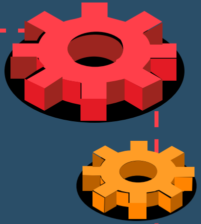
- Logistic regression is a statistical method used for binary classification problems.
- The model uses a linear combination of input features, which is then transformed using the logistic function to produce the probability of na event

XGBoost

- XGBoost, short for Extreme Gradient Boosting, is an optimized implementation of the gradient boosting algorithm. It is a highly efficient and scalable machine learning algorithm known for its exceptional performance in a wide range of tasks, including regression, classification, and ranking problems.

Gradient Boosting

- Gradient Boosting is a powerful machine learning algorithm that belongs to the family of ensemble methods, which combines multiple weak learners to create a strong learner.



Developed Models

Neural Networks

- Layers of interconnected nodes, or neurons, that perform mathematical operations on the input data to produce na output. Deep neural networks with many layers have been shown to be particularly effective for complex tasks such as image and speech recognition

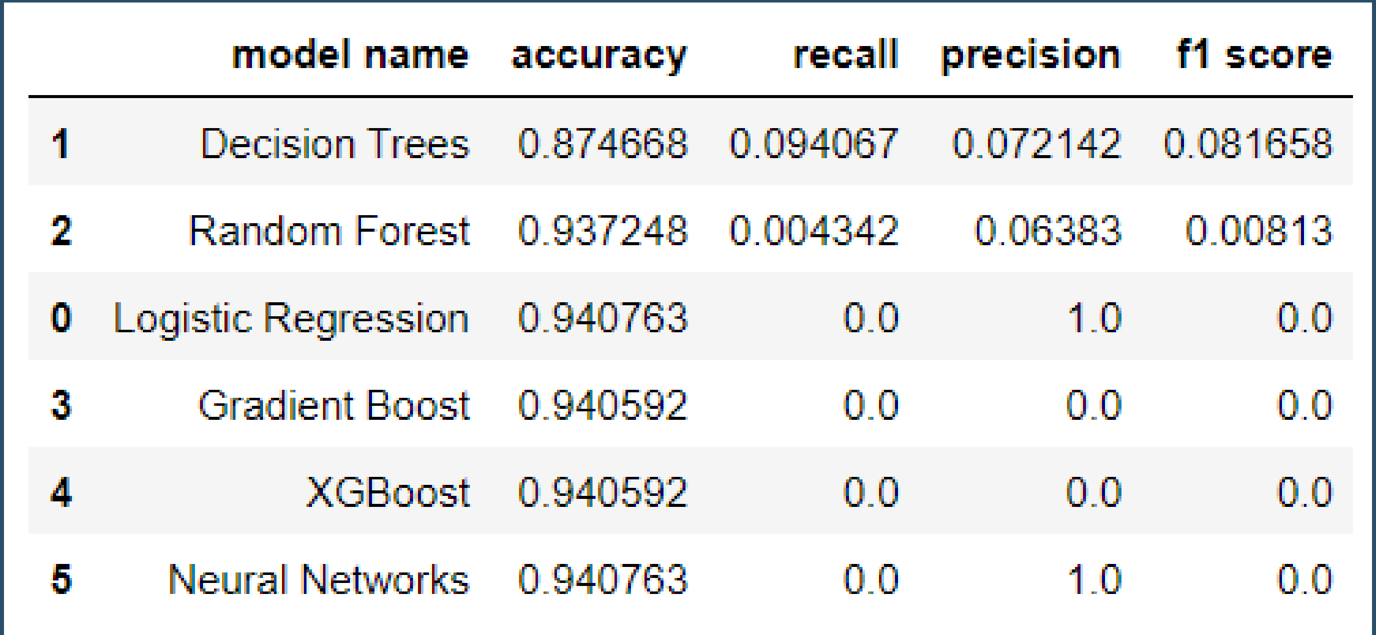
Decision Trees

- Each internal node of the tree represents a decision based on a feature value, and each leaf node represents a prediction. The tree is constructed recursively by choosing the feature that best splits the data at each node.

Random Forests

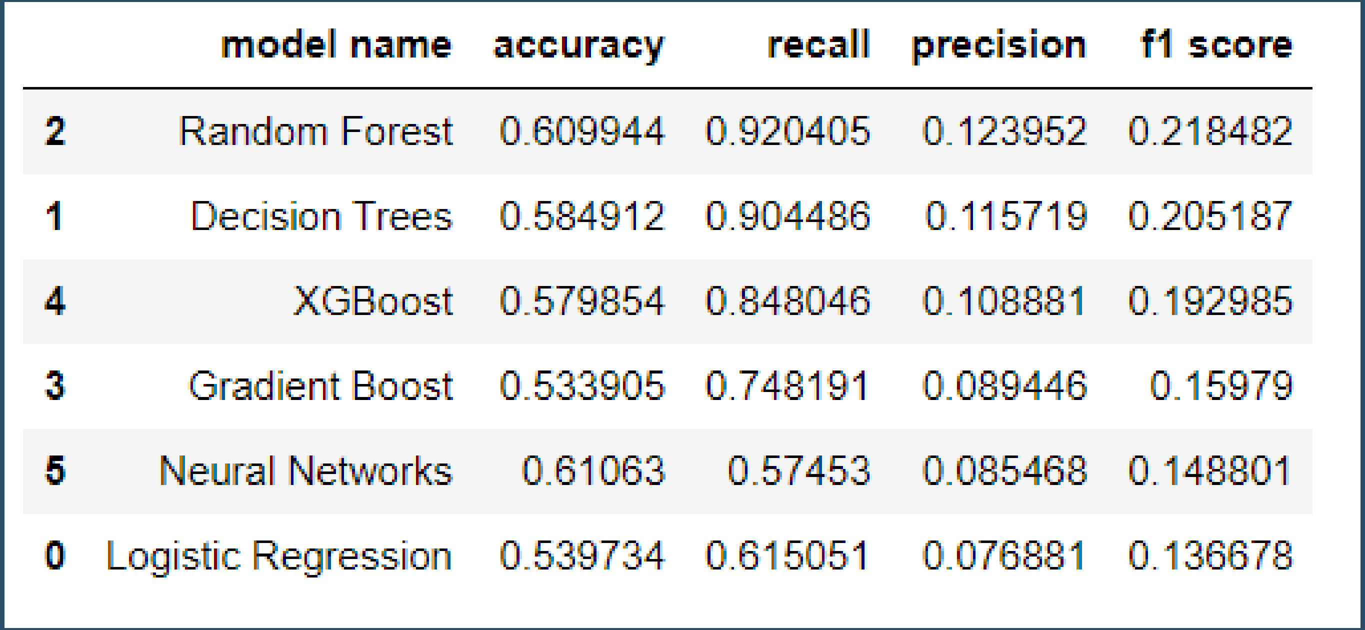
- Each tree in the forest is trained on a ranom subset of the data and a random subset of feautres, ensuring diversity in the models and reducing overfitting.
- The final prediction is then made by aggregating the predictions of all the individual trees.

Comparison – original data



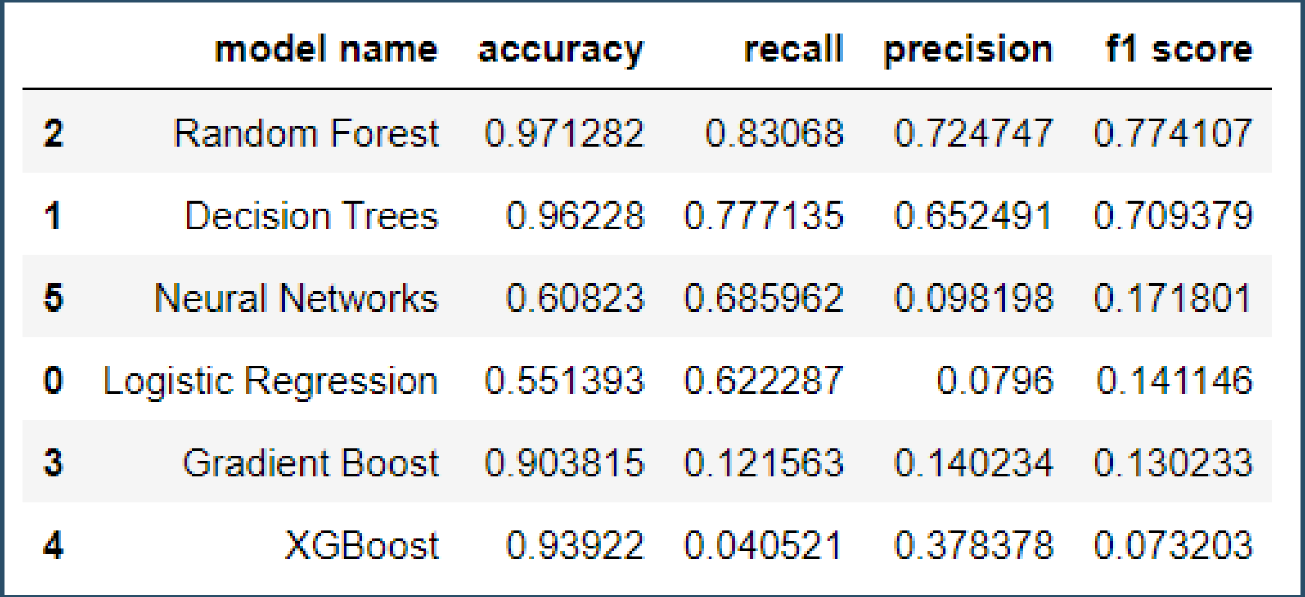
	model name	accuracy	recall	precision	f1 score
1	Decision Trees	0.874668	0.094067	0.072142	0.081658
2	Random Forest	0.937248	0.004342	0.06383	0.00813
0	Logistic Regression	0.940763	0.0	1.0	0.0
3	Gradient Boost	0.940592	0.0	0.0	0.0
4	XGBoost	0.940592	0.0	0.0	0.0
5	Neural Networks	0.940763	0.0	1.0	0.0

Comparison – undersample



	model name	accuracy	recall	precision	f1 score
2	Random Forest	0.609944	0.920405	0.123952	0.218482
1	Decision Trees	0.584912	0.904486	0.115719	0.205187
4	XGBoost	0.579854	0.848046	0.108881	0.192985
3	Gradient Boost	0.533905	0.748191	0.089446	0.15979
5	Neural Networks	0.61063	0.57453	0.085468	0.148801
0	Logistic Regression	0.539734	0.615051	0.076881	0.136678

Comparison – oversample



	model name	accuracy	recall	precision	f1 score
2	Random Forest	0.971282	0.83068	0.724747	0.774107
1	Decision Trees	0.96228	0.777135	0.652491	0.709379
5	Neural Networks	0.60823	0.685962	0.098198	0.171801
0	Logistic Regression	0.551393	0.622287	0.0796	0.141146
3	Gradient Boost	0.903815	0.121563	0.140234	0.130233
4	XGBoost	0.93922	0.040521	0.378378	0.073203