Data Processing:

First, we divide the data into three types. One type is binary data with 'yes' or 'no' responses.
The second type is numeric data, which has undergone normalization.
The third type is categorical data with multiple categories.
For age data containing years and months, we determine the actual age by calculating the total number of months.
For power and torque, we simply take the values from the front as the data.
Model Selection:
From a machine learning perspective, we focus on solving problems using mathematical approaches rather than expanding the feature dimension with deep learning, which largely requires us to address issues from practical situations. In terms of model selection, both for regression and classification, we used Random Forest as the basic concept to address these issues, which yielded relatively good results. XGBoost could improve our results to some extent, but it was not as fast as Random Forest, hence we chose RF.

Fine Tuning:
We start by calculating the minimum and maximum values to find the desired regression interval.
We found that the train data distribution ranges from 23 to 78 years old,
while the test data distribution ranges from 23 to 80 years old.
Considering the practical application scenario, typically, the minimum driving age is 18 years, and the maximum is around 80 years.
With data from 40,000 individuals, we realized that using pure regression to address this issue, regression could not accurately predict ages below 23 or above 47. This is related to the normal distribution of the data. Both train and test data have an average age of around 37 years, with a standard deviation of about 9.8. Therefore, traditional regression better fits within the age interval of [28, 47]. For this mathematical phenomenon, we subtracted 23 from everyone's age and treated the traditional regression problem as a classification problem. This adjustment allowed the prediction interval to be expanded to ages 23-70, addressing the issue of excessive prediction MSE.

The second issue is the classic imbalanced data problem, where using basic random forest resulted in severe overfitting. Therefore, we calculated the weight ratio of 1s and 0s in the train and test data, which was around 15:1. However, based on our results from setting different weights, we found that a 10:1 ratio significantly improved the model’s overfitting issue.
Additionally, for the single-category issue, using around 40 features was too many. So, we used feature importance to identify the most critical 10 features as our final selection. This approach

also greatly alleviated the overfitting problem.

feature_importances_sorted.head(10)['feature']	Test Macro F1-Score: 0.5348733730163325
feature_importances_sorted.head(20)['feature']	Test Macro F1-Score: 0.5316265446628398
feature_importances_sorted.head(30)['feature']	Test Macro F1-Score: 0.5303338690398548
feature_importances_sorted.head(40)['feature']	Test Macro F1-Score: 0.5297756702753346

RandomForestClassifier

N_estimators	Max_depth	Test MSE
70	None	8.245
70	3	121.685
70	5	89.693
70	7	65.9482
70	9	47.5454
70	11	41.3714
70	20	13.1978
100	None	7.258
100	3	138.3192
100	5	83.6036
100	7	59.986
100	9	48.7552
100	11	40.0628
100	20	12.5554
200	None	8.0066
200	3	115.3266
200	5	81.6046
200	7	62.751
200	9	48.8548
200	11	40.0702
200	20	11.307

Business value:

Enhanced Accuracy in Age Prediction:

By focusing on calculating actual age and adjusting the prediction interval, this model provides more accurate age predictions for policyholders. This accuracy is crucial for tailoring insurance products and pricing according to age-related risk factors. Insurance premiums and coverage options can be more precisely aligned with the specific risk profiles associated with different age groups, potentially increasing the attractiveness of insurance offers to customers.

Improved Claim Prediction:

This classification model predicts the likelihood of a policyholder lodging a claim within the next six

months. This is immensely valuable for risk management and financial planning within an insurance company. By identifying individuals with a higher probability of filing claims, insurers can adjust premiums, set aside appropriate reserves for expected claims, and enhance their customer service approaches to manage high-risk clients proactively.