HCS502: Assessment - AI/ML Project - 2310190

Thomas Morton

June 2025

Abstract

This study aims to explore and employ machine learning techniques in order to predict the cost of insurance premiums for an individual based on details about the driver as well as information about the vehicle. The research will also outline how each of these features affects the insurance premium for a driver. The data set being used has 1000 records containing the aforementioned data. This dataset will be used to train and analyse both simple and multivariate regression models, as well as clustering algorithms. Data preprocessing was carried out on the data, which consisted of normalisation in the form of scaling and anomaly identification.

1 Introduction

Car insurance and its workings are a mystery to many. It traditionally relies on analysis and experts in the domain of car insurance to make assumption and asses the risk of certain drivers and vehicles and set a driver's insurance premium based on this knowledge and experience. In the modern world, with comparison sites and insurers in the 100's or even 1000's, artificial intelligence (AI), along with machine learning (ML), probably already run the business side of most insurers on the market. Although it may seem unfair to let these algorithms control the prices, they use data-driven methods which is trained on raw facts from the world of motoring. This technology offers an automated solution, saving the companies a lot of money in training individual staff members on the risk factors to look out for in a person.

This project investigates how effective a simple regression algorithm, using one factor, or a multivariate algorithm, using multiple factors, can be used to predict car insurance premiums based on the information provided. Linear regression is known for its interpretability and efficiency in a computational sense [**scikit˙LR**], which allows for easy identification of trends and displaying the correlation between key factors. In applying these models to a structured dataset, this study aims to explore how effectively insurance premiums can be predicted, which will allow the assessment of the model accuracy and evaluate how this can also present the risk of over-fitting to the data.

Understanding the relationship between these factors is imperative to the research into the workings of insurance company pricing structures. This research seeks to provide more transparent pricing and show the factors that have the biggest impact on an insurance policy.

2 Dataset and Pre-processing

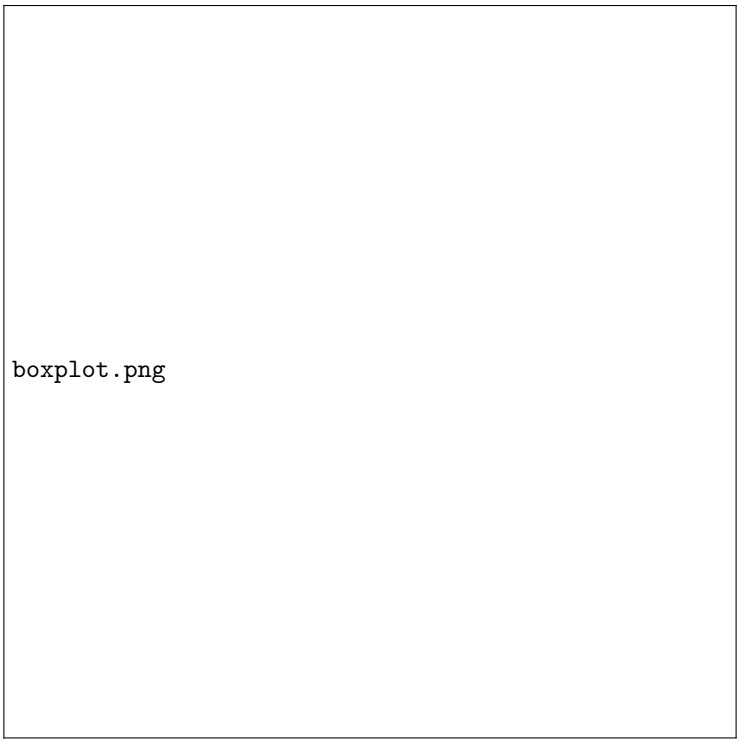

Figure 1: Boxplot of Normalised Variables

3 Exploratory Data Analysis (EDA)

Figure 2: Correlation Heatmap of All Features

4 Modelling and Evaluation

Linear regression models aim to establish a linear relationship between input features x_i and the continuous target variable y (insurance premium), typically expressed as:

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

where β_0 is the intercept and β_i are the coefficients learned from data [**hastie˙09˙elements-of.statistical-learning**].

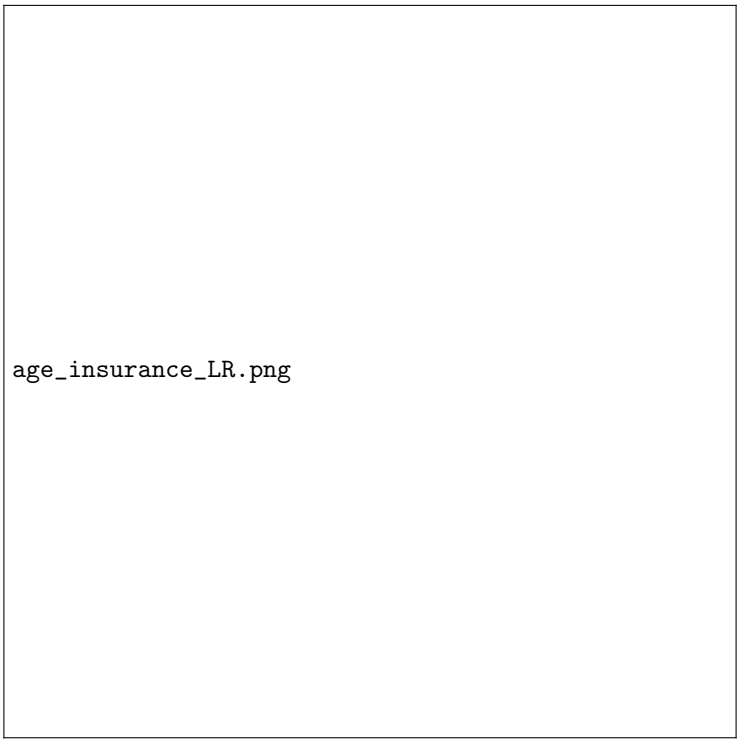

Figure 3: Regression Line: Driver Age vs Insurance Premium

4.1 Model Comparison

- Simple linear regression provides interpretability but lower accuracy.
- Multivariate regression achieves high performance on synthetic data but raises questions about generalizability.
- Further work may include Ridge/Lasso regression to assess generalisation.

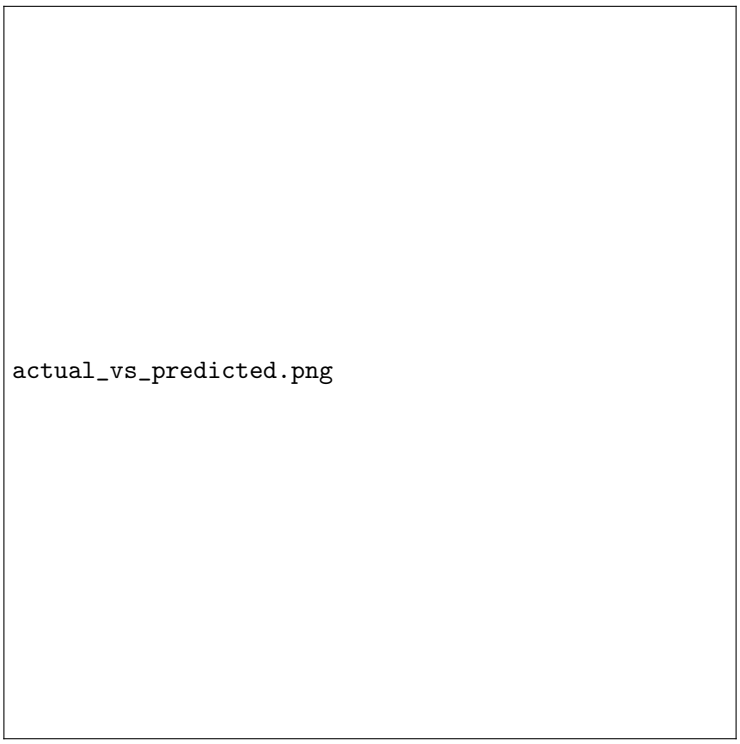

Figure 4: Actual vs Predicted Premiums (Multivariate Model)

5 Software Environment

6 Software Environment

The machine learning pipeline was implemented in Python 3.13.2 using popular scientific libraries including `pandas` for data manipulation, `scikit-learn` for modeling, and `matplotlib` and `seaborn` for visualization.

7 Results Analysis

8 Ethical Considerations

The use of machine learning for predicting car insurance premiums introduces several ethical concerns:

Fairness and Bias:

Data Privacy:

Transparency:

Accountability:

9 Conclusion