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, clustering algorithms and neural networks. 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[2], 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

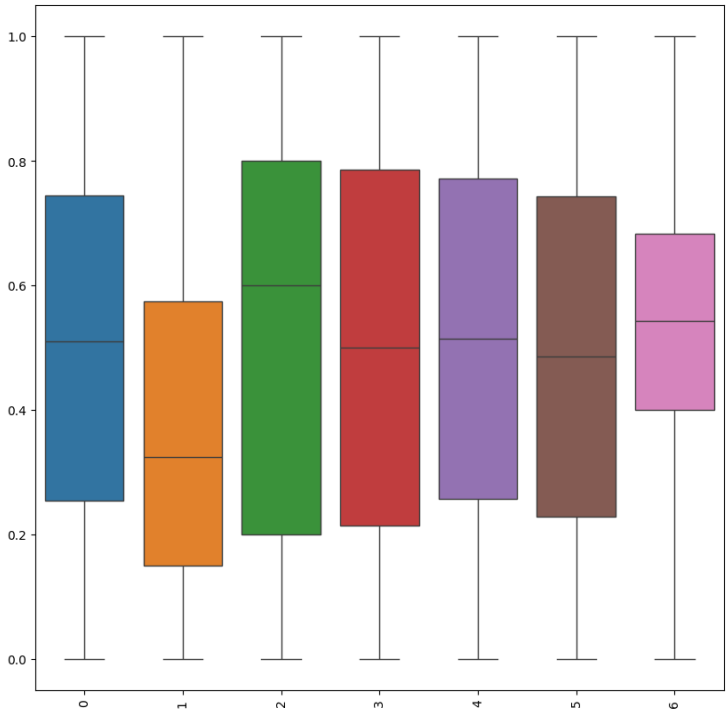

Figure 1: Boxplot of Normalised Variables

The dataset was sources from Kaggle.com and provides an array of information including the details about a driver, their vehicle and the insurance premium they were quoted because of this[4]. The dataset consists of 1000 rows of data containing numerical features along with the insurance price. Thee features include; Driver Age, Driver Experience, Previous Accidents, Annual Mileage (KM), Car Manufacturing Year, Car Age and finally, the taget variable, Insurance Premium in US dollars.

To ensure the data was of a high enough quality, there was some initial exploratory analysis that took place in the form of normalisation, box plots and statistical summary. The inital check was to make sure where were no null values in the data, of which there are none as shown below in the output of an 'isnull().sum()' check. SKlearn's standard scaler [3] was used to normalise the data for the boxplots which makes it a lot easier to plot and pick out any outliers, of which there are none in this instance. Due to the nature of our data, all of the variables are integers or floats, this was confirmed by using 'df.dtypes' against the imported dataframe.

```
Driver Age                    0
```

```
Driver Experience            0
Previous Accidents           0
Annual Mileage (x1000 km)    0
Car Manufacturing Year       0
Car Age                      0
Insurance Premium ($)        0
dtype: int64
```

3 Exploratory Data Analysis (EDA)

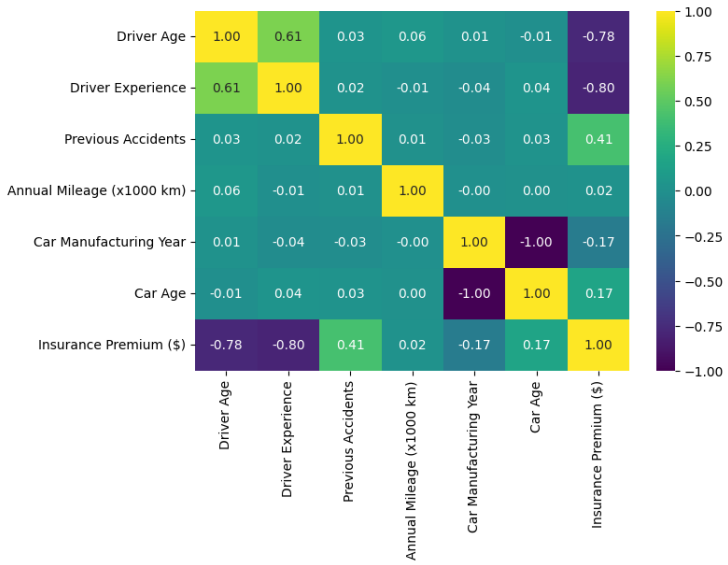

Figure 2: Correlation Heatmap of All Features

Using seaborn makes it simple to view a heatmap of the correlation between all of the variables. This is shown in the above figure comparing the relationip of any of the features against any other feature. In this heatmap it points out the strong negative correlations between 'Driver Age' and 'Driver Experience', these show a negative corealtion of -0.78 and -0.80 respectively. This suggests that older and more expericenced drivers are considered safer drivers which in turn allows them to have a lower insurance premium. This alligns with general risk assesment principles that expericenced drivers tend to be safer and older drivers often more sensible on the road.

As expected, previous accidents seems to have a fairly strong positive correlation against insurance premiums. This implies that the more accidents that a driver has, the more at risk they are on the road. This again aligns with general risk assessment principles that drivers that have accidents more often, are more likely to have another accident in future. Other features in our dataset showed weaker and more negligable correlations, meaning that the correlation is not as strong and these have less of an effect on the price of car insurance. It would be interesting to see if other factors were added to this dataset, how much effect they would have. Such as engine size (cc), engine power (bhp), driveterrain layout (AWD, FWD, RWD), however those are not available for this dataset.

4 Modelling and Evaluation

Choosing a model to use for the application requires some testing of various models and evaluating the results. Some obvious choices are simple linear regression and multivariate regression. K-means and K-nearest neighbor are also useful for seperating data into different groups, however for predicting a continuous value such as insurance premiums, linear regression is the more appropirate. Dispite this, this section will cover all aforementioned to ensure no opportunities for appropirate models are missed.

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[1].

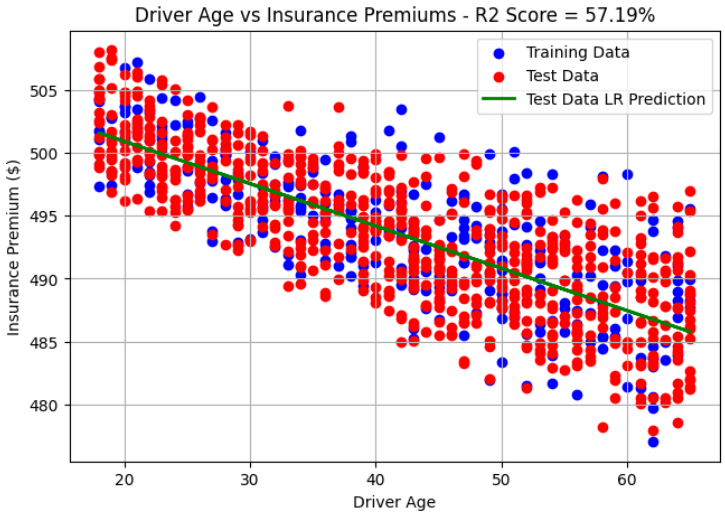

Figure 3: Regression Line: Driver Age vs Insurance Premium

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.

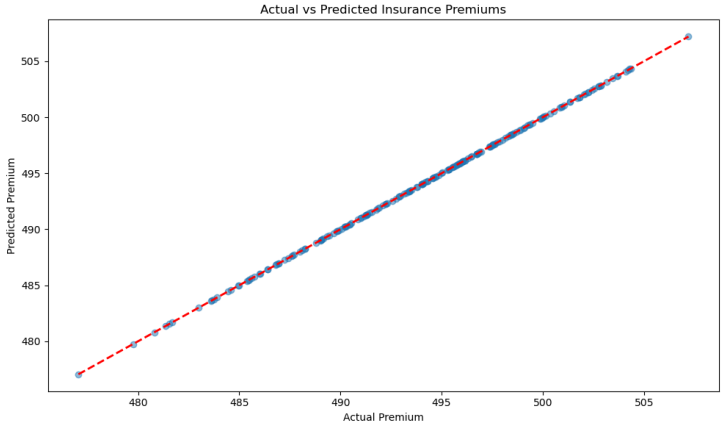

Figure 4: Actual vs Predicted Premiums (Multivariate Model)

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

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[2] *Scikit learn: Linear Regression*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression.

[3] *Scikit learn: Standard Scaler*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler.

[4] Govindaram Sriram. *Car Insurance Premium Dataset*. Jan. 2025. URL: https://www.kaggle.com/datasets/govindaramsriram/car-insurance-premium-dataset.