Data-Driven Insurance Premium Prediction

Amy Hatman

DSC680 – Applied Data Science

Business Problem

Insurance companies must strike a delicate balance in pricing premiums. If premiums are too high, customers will seek competitors, reducing market share. If premiums are too low, insurers risk financial losses when claims exceed collected revenue. Traditional actuarial methods, which rely heavily on broad averages and demographic categories, provide a foundation but fail to capture the full complexity of modern driver behavior, vehicle differences, and risk profiles (Insurance Information Institute, 2023).

Background and History

The auto insurance industry has historically used actuarial tables and statistical averages to guide premium pricing. While this approach has been successful in establishing a baseline for risk, it oversimplifies the nuances of individual drivers and vehicles. Advances in data availability and computational power now allow insurers to model risk with far greater precision (Deloitte, 2021). By incorporating driver demographics, driving history, and vehicle information, insurers can refine premium estimates and reduce reliance on broad categories. This transition from traditional methods to data-driven modeling parallels larger trends across financial services, where predictive analytics and machine learning are increasingly used to optimize decision-making (Abbott, 2014).

Data Explanation

The dataset used for this analysis contains 1,000 driver. Each record pairs a driver's attributes with their insurance premium. Features include driver age, driving experience, number of previous accidents, annual mileage, car manufacturing year, car age, and the target variable: insurance premium in US dollars (Sriram, n.d.).

Data preparation steps included confirming there were no missing values, validating ranges such as ensuring driver age fell between eighteen and sixty-five, and addressing redundancy. Car age and car manufacturing year were perfectly correlated, so only one was needed for analysis. All data were numeric, simplifying preprocessing.

Methods

The analysis began with exploratory data analysis (EDA). Histograms and boxplots were used to study distributions and identify outliers. A correlation matrix and heatmap revealed relationships between predictors and premiums. Modeling approaches build upon prior work in predictive analytics (Abbott, 2014).

Analysis

Premiums were strongly negatively correlated with driver age and driving experience, confirming that younger, less experienced drivers pay higher premiums. Premiums were moderately positively correlated with accident history. These results are consistent with industry research documenting the importance of these variables in shaping risk profiles (Insurance Information Institute, 2023).

Assumptions

This project assumes the dataset is representative of real-world driver populations. It also assumes accurate recording of accidents, mileage, and other driver attributes. The relationships observed here are taken as reflective of actual insurance risk, though in reality many more variables may be involved.

Limitations

The dataset is limited in scope, with only 1,000 records. Premiums also fall within a narrow range, which may mask variability present in real-world insurance markets. Geographic, demographic, and behavioral factors such as location, gender, or telematics data were not included, reducing predictive power.

Challenges

The project faced several challenges. Feature overlap, particularly between driver age and experience, introduced potential multicollinearity. Ensuring fairness while including age as a predictor raised ethical concerns. The small dataset limited the generalizability of findings. Finally, the trade-off between interpretability and predictive power posed a challenge in model selection.

Future Uses and Applications

Future extensions of this project could include telematics data such as speed, acceleration, and time-of-day driving patterns. Regional and demographic attributes would enable localized pricing models. With larger datasets, real-time models could support usage-based insurance, where premiums adjust dynamically based on driver behavior.

Recommendations

Insurers should focus on driver-related factors in premium pricing while expanding their data sources to capture richer behavioral information. Regular recalibration of models is necessary to ensure fairness and accuracy. Investments in telematics will provide valuable insights into risk beyond traditional demographics.

Implementation Plan

The plan includes developing predictive models using regression and machine learning techniques, validating with historical claims data, and gradually integrating models into production systems. Pilot programs can be launched with select customer segments, followed by broader rollout once results are validated. Ongoing monitoring and auditing are critical for ensuring fairness and compliance.

Ethical Assessment

As Deloitte (2021) highlights, advanced analytics create opportunities but also ethical challenges. Fairness is critical, since variables such as age may disadvantage specific groups. Transparency is also essential: customers should be able to understand in general terms why they are charged a given premium. These principles align with best practices in predictive analytics (Abbott, 2014).

Conclusion

The analysis suggests that driver-related characteristics are the most significant predictors of insurance premiums. Age, experience, and accident history together explain much of the variation in premiums, while car-related attributes and mileage have limited impact. These insights reinforce the industry view that driver behavior and history are the dominant indicators of insurance risk.

Illustrations

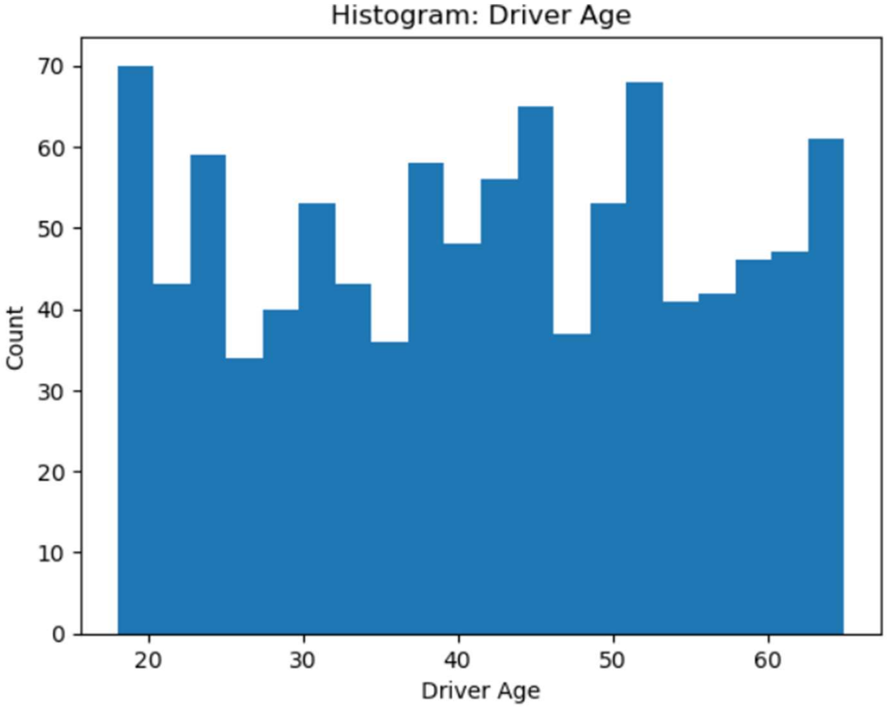


Figure 1: Histogram of driver age distribution.

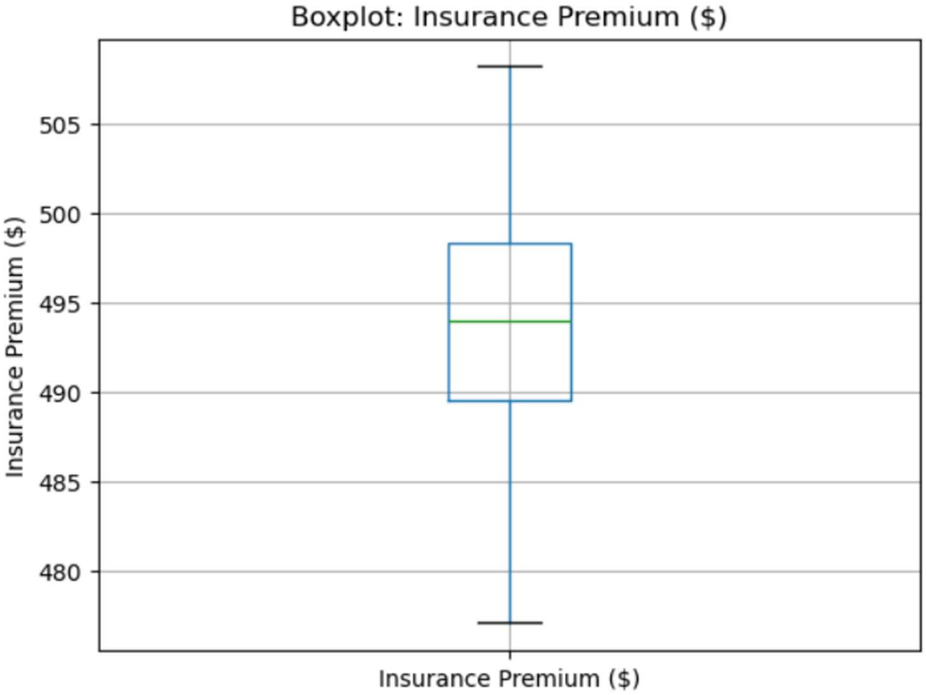


Figure 2: Boxplot of insurance premiums.

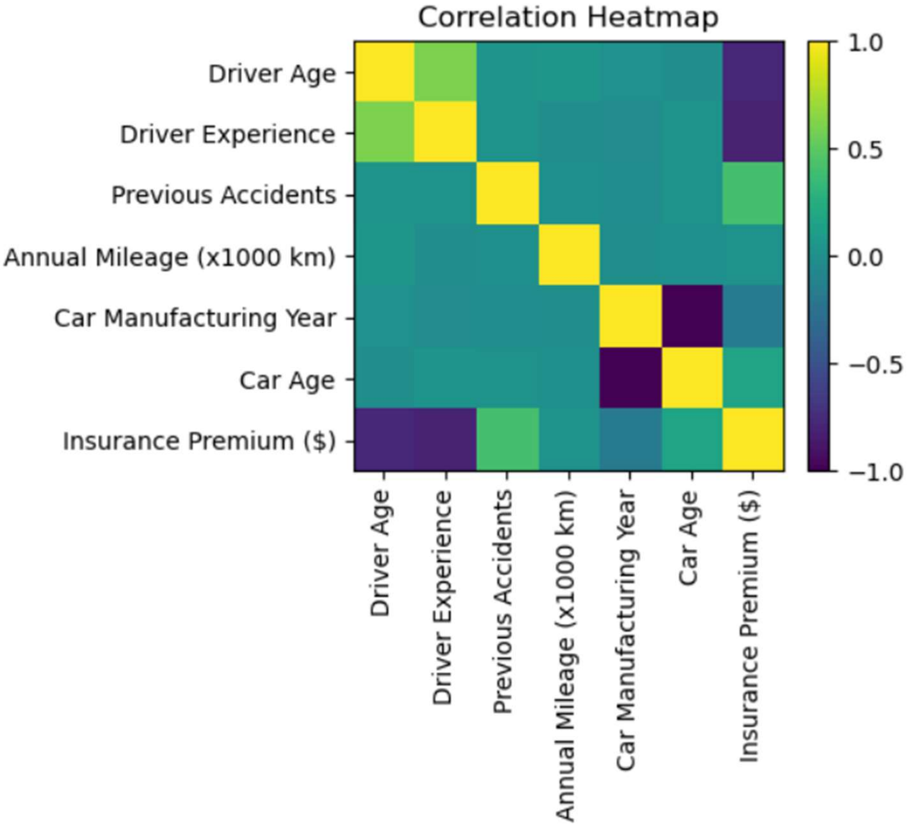


Figure 3: Correlation heatmap of features.

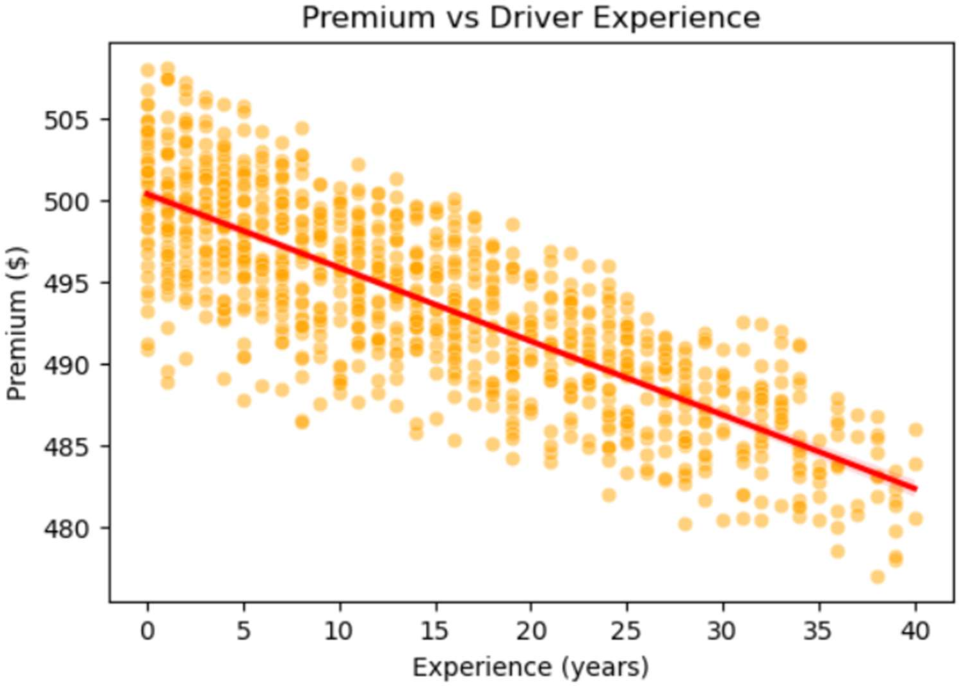


Figure 4: Scatterplot of insurance premium versus driving experience with regression line.

Appendix: Supporting Documentation

Data Dictionary

Driver Age: Age of the driver in years.

Driver Experience: Years of driving experience.

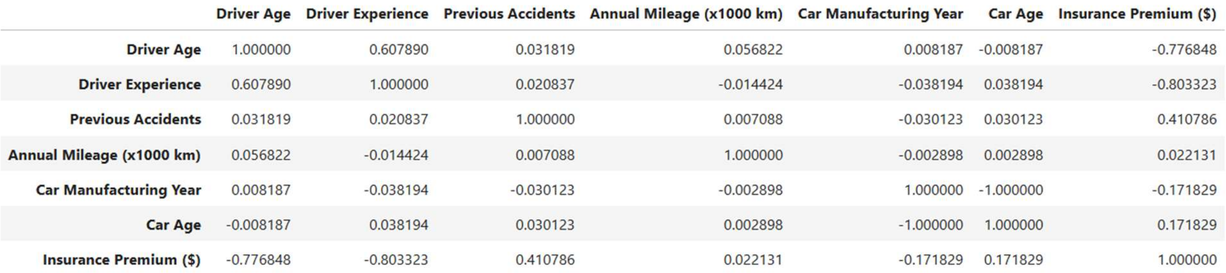
Previous Accidents: Number of past accidents recorded.

Annual Mileage: Yearly driving distance in thousands of kilometers.

Car Age: Age of the vehicle in years.

Insurance Premium: Annual cost of insurance in US dollars.

Summary Statistics



	Driver Age	Driver Experience	Previous Accidents	Annual Mileage (x1000 km)	Car Manufacturing Year	Car Age	Insurance Premium ($)
Driver Age	1.000000	0.607890	0.031819	0.056822	0.008187	-0.008187	-0.776848
Driver Experience	0.607890	1.000000	0.020837	-0.014424	-0.038194	0.038194	-0.803323
Previous Accidents	0.031819	0.020837	1.000000	0.007088	-0.030123	0.030123	0.410786
Annual Mileage (x1000 km)	0.056822	-0.014424	0.007088	1.000000	-0.002898	0.002898	0.022131
Car Manufacturing Year	0.008187	-0.038194	-0.030123	-0.002898	1.000000	-1.000000	-0.171829
Car Age	-0.008187	0.038194	0.030123	0.002898	-1.000000	1.000000	0.171829
Insurance Premium ($)	-0.776848	-0.803323	0.410786	0.022131	-0.171829	0.171829	1.000000

Data Preparation Notes

The dataset was checked for missing values, and none were found. Variables were numeric and did not require encoding. Redundant features such as car age and manufacturing year were reconciled. Distributions were validated against expected ranges for driver demographics and vehicle characteristics.

References

Abbott, D. (2014). Applied predictive analytics: Principles and techniques for the professional data analyst. Wiley.

Deloitte. (2021). Insurance industry outlook: Balancing risk, regulation, and innovation. Deloitte Insights.

Insurance Information Institute. (2023). Auto insurance: Facts and statistics. Insurance Information Institute.

Sriram, G. (n.d.). Car insurance premium dataset. Kaggle.