2024

# Recommended Car Insurance Premiums for Low-Risk Customers

Julie Campbell

# Table of Contents

## Business Problem

Remaining competitive in the insurance industry requires providing customers with a price that they are willing to pay while also mitigating the risk of accidents. This often means providing personal pricing according to the risk that the insurer is willing to take on the from the customer. The research will be to determine which features of a person's insurance policy have the greatest impact on premium based on past issuances. A machine learning model will be built to estimate a competitive premium for new clients.

## Background

The ability to recommend a car insurance premium for customers will reduce the agent's time spent on low-risk policies. This gives the agent more time to focus on the medium to high-risk policies instead that would have a greater impact on the insurance company. The dataset collected came from the Ethiopian Insurance Corporation and contained a sample of insurance policies from 2014-2018 for vehicle insurance. For this report's purposes, low-risk customers will be defined as those with vehicles that have no insured value (liability coverage) and expect a relatively low premium defined as less than $6,000. Due to the complexity of how different vehicles are utilized and potentially damaged only consumer vehicles (automobile, truck, trailer) will be considered.

## Data Preparation

During preparation, multiple fields had to be converted to integer or date types once imported. Due to the low number of null values of the target value, premium, the mean was used to replace null values. Several filters had to be implemented to categorize data:

- Timeframe: January 2018 or later
- Vehicles less than or equal to 10 years old
- Small or Medium vehicles size
- Vehicle makes with at least 100 samples
- No claims paid

There was also a consolidation of vehicle make names due to spelling errors and to the usage categories for overlapping or not relevant values.

## Methods

The premium value was documented as a continuous value of less than $6,000. A linear model should be created and evaluated to predict the premium price that should be charged to the low-risk customer. The correlation between the values was compared across all fields and to premium prices themselves. Features will be reduced based on the p-value once a model has been created. Several linear regression model types will be considered including standard linear, ridge, and lasso.

## Analysis

Exploration showed that most premium values were documented as less than $1,000 with some outliers between $4,000-5,000. This only includes the records as defined as low-risk customers. The relationship between premium amount and sex was compared on Figure 1.

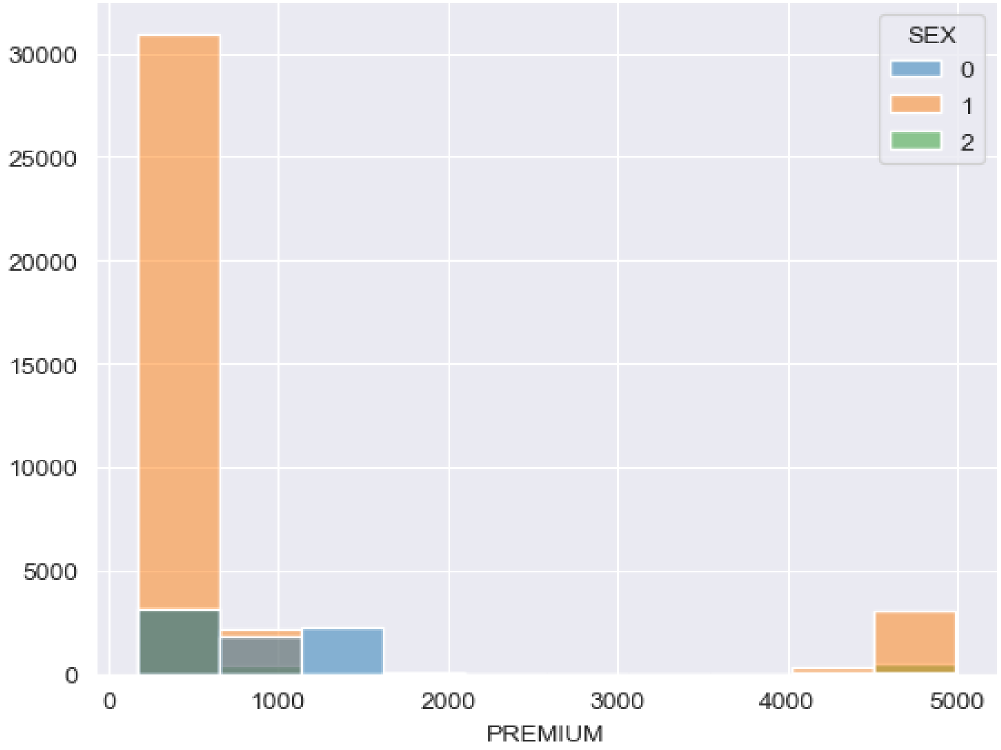


Figure 1: Premium Amount by Sex

The production year of the vehicles is concentrated between 2012-2018. It was also noted that motorcycles and taxi usage made up most of the vehicle policies.

| Types | Correlation |
|---|---|
| Type: Truck | 0.9711 |
| Make: Isuzu | 0.9572 |
| CCM Ton | 0.8624 |
| Usage: Commercial | 0.5561 |
| Type: Pick-up | 0.0744 |

Table 1: Top 5 Positive Correlation Features

| Types | Correlation |
|---|---|
| Type: Motorcycle | -0.7149 |
| Usage: Taxi | -0.4016 |
| Make: Bajaji | -0.3714 |
| Usage: Private | -0.3714 |
| Production Year | -0.1597 |

Table 2: Top 5 Negative Correlation Features

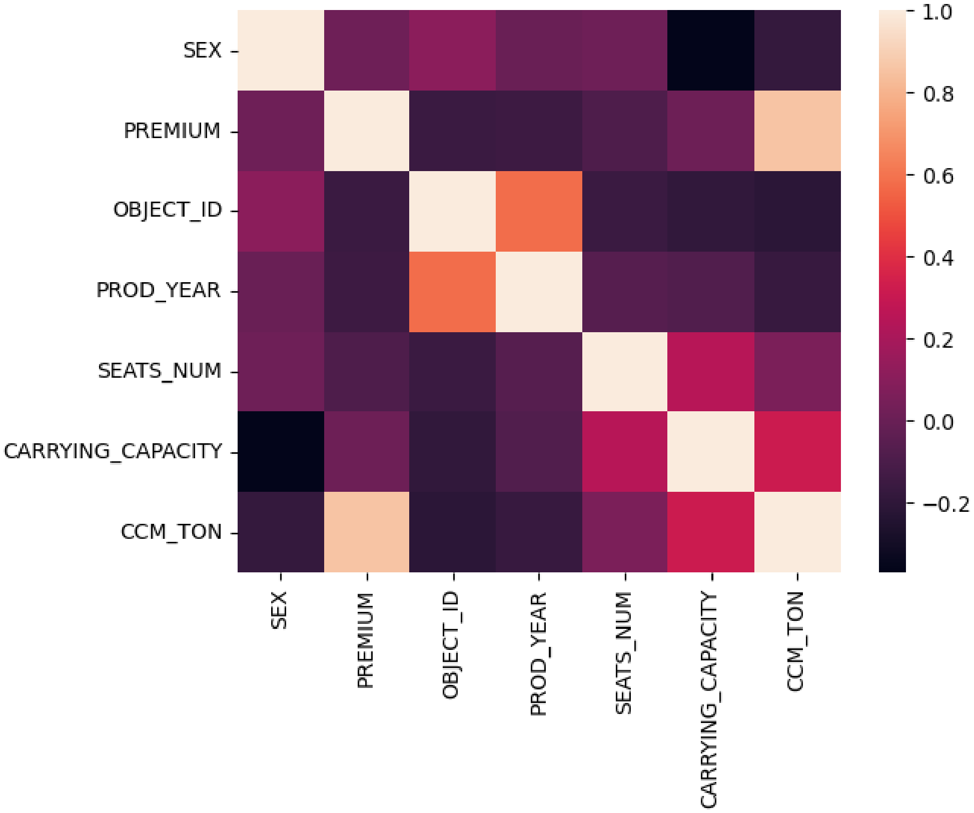


Figure 2: Correlation Matrix

The prepared dataset was split into 70/30 for model training. The features were scaled by the standard scalar. After comparing the p-value, the following features were chosen for the model:

- Sex
- Carrying Capacity
- CCM Ton
- Seat Number
- Vehicle Type: Automobile, motorcycle, pick-up, station wagon
- Make: Bajaji, Nissan, Isuzu
- Usage: Taxi

Several linear regression models were considered:

| Types | R2 Score | RMSE |
|---|---|---|
| Linear | 98.201% | $155.86 |
| Ridge | 98.201% | $155.87 |
| Lasso | 98.197% | $156.05 |

Table 3: SARIMAX Model RMSE

## Conclusion

The default linear regression model performed the best with a 98.201% R2 score and $155.86 RMSE. With the premium concentrated between $0-2000, the RMSE is an acceptable amount with the tradeoff of saving the insurance agent’s time. In Figure 3, the residuals between the actual and predicted values were plotted to determine the model's success.

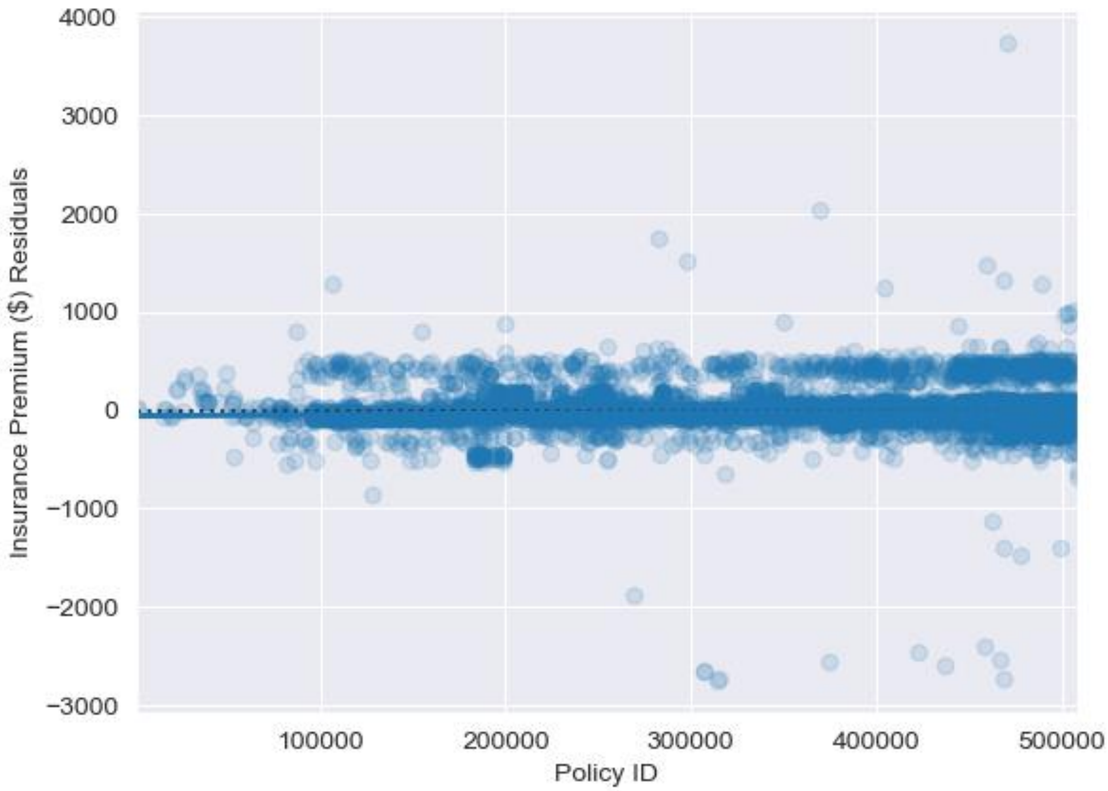


Figure 3: Linear Regression Model Correlation

## Assumptions

The primary assumption that was made during research is that most of the dataset contained 0 for the insured value which indicated that it is likely associated with a liability policy. A liability policy is defined as a product that provides protection against claims resulting from injuries and damage to other people or property (Kagan, 2022). To find the criteria for low-risk customers, the premium limit was chosen based on the unfiltered histogram of premium after all other filters were applied.

### Limitations and Challenges

With the dataset filtered for those who have no insured value and relatively low premium, it is limited to customers that fit that criteria purposefully so that insurance agents can focus on more complex cases. The model should not replace agent validation time once it enters underwriting.

## Future Applications

The basis of this model should be used by insurance companies to reduce the workload on their agents by using the linear regression model to create an initial premium recommendation.

### Recommendations

It is recommended to validate the assumptions made above and continue to train the model based on more recent data.

### Implementation Plan

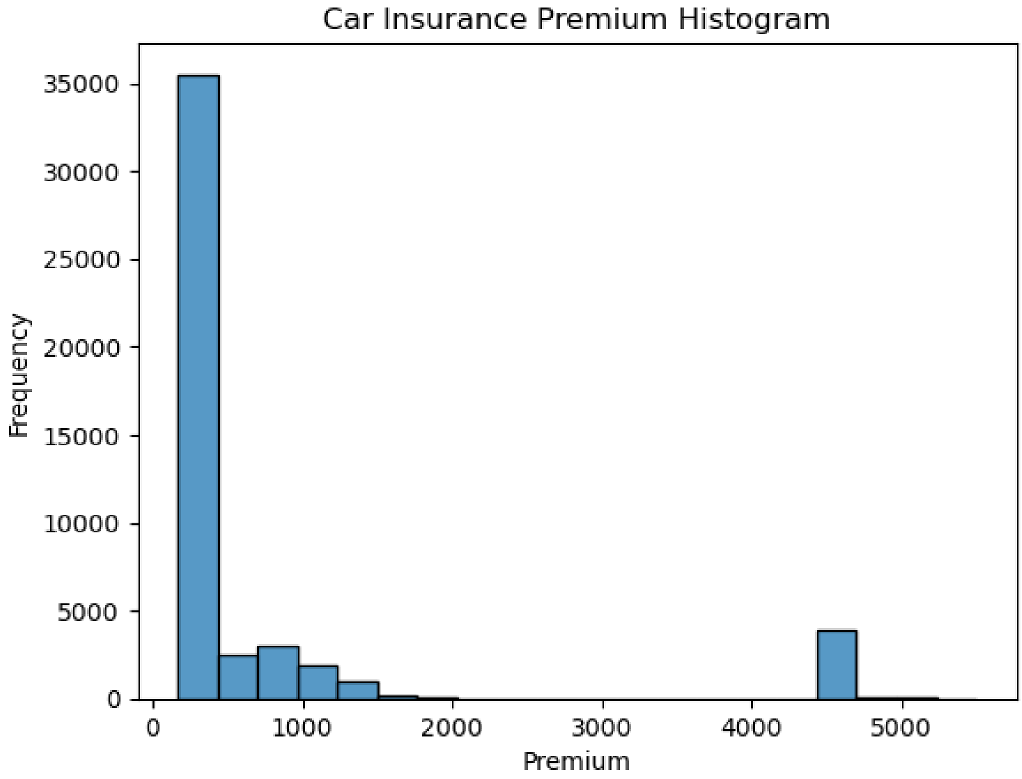
The example dataset and model will be posted on GitHub for open-source use for employees of insurance companies to evaluate and determine if there is a similar pattern that could be used on their own dataset.

## Ethical Assessment

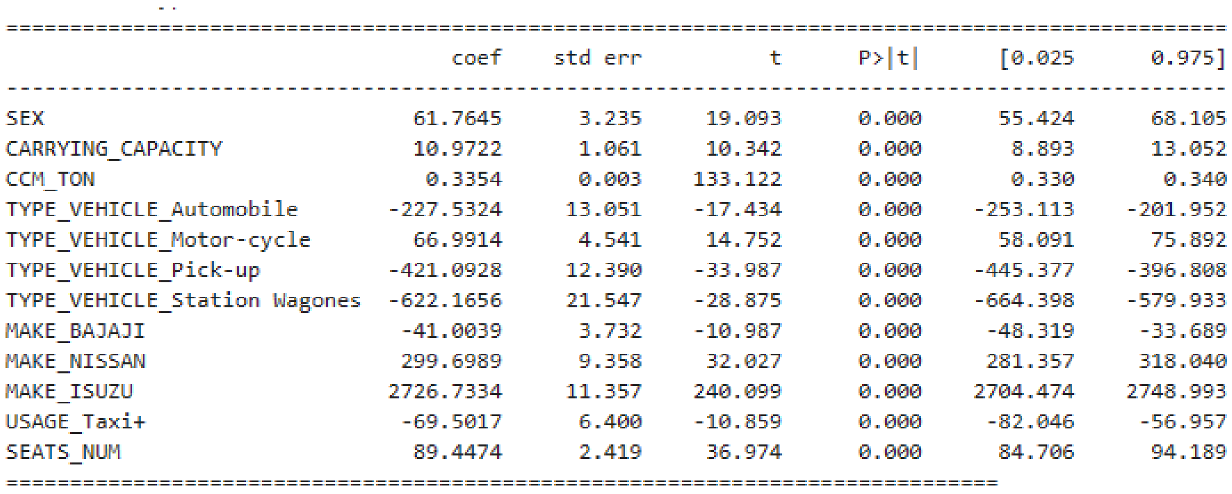
The dataset was explained to be received from Ethiopian Insurance Corporation with personal data removed, but it contains data over 6 years old that may or may not be accurate or approved for distribution outside of an educational setting.

# Appendix

Appendix A: Car Insurance Premium Histogram



Appendix B: p-value from OLS Regression Results



```
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
SEX                           61.7645      3.235     19.093      0.000      55.424      68.105
CARRYING_CAPACITY             10.9722      1.061     10.342      0.000       8.893      13.052
CCM_TON                        0.3354      0.003    133.122      0.000       0.330       0.340
TYPE_VEHICLE_Automobile     -227.5324     13.051    -17.434      0.000    -253.113    -201.952
TYPE_VEHICLE_Motor-cycle      66.9914      4.541     14.752      0.000      58.091      75.892
TYPE_VEHICLE_Pick-up        -421.0928     12.390    -33.987      0.000    -445.377    -396.808
TYPE_VEHICLE_Station Wagones -622.1656     21.547    -28.875      0.000    -664.398    -579.933
MAKE_BAJAJI                  -41.0039      3.732    -10.987      0.000     -48.319     -33.689
MAKE_NISSAN                  299.6989      9.358     32.027      0.000     281.357     318.040
MAKE_ISUZU                  2726.7334     11.357    240.099      0.000    2704.474    2748.993
USAGE_Taxi+                  -69.5017      6.400    -10.859      0.000     -82.046     -56.957
SEATS_NUM                     89.4474      2.419     36.974      0.000      84.706      94.189
==============================================================================
```

## References

Kagan, J. (2022, June 21). *Liability Insurance: What It Is, How It Works, Major Types*. Retrieved from Investopedia: https://www.investopedia.com/terms/l/liability_insurance.asp

Momeni, M. (2023). *Vehicle Insurance Data*. Retrieved from Kaggle: https://www.kaggle.com/datasets/imtkaggleteam/vehicle-insurance-data?select=motor_data14-2018.csv