Understanding Car Insurance Premiums

Goals:

1. The purpose of this project was to determine the most important features in this data set that predicted whether or not the customer would have high insurance premiums.
2. To use machine learning to determine the demographic of people most likely to have high premiums.

Introduction

The subject area for this project is the car insurance industry. This is a very relatable topic for many people and applying data science techniques to understand this industry is something that many people can relate to.

This dataset was acquired from Emcien, a data science solution software company. The original dataset had 9134 rows and 26 features. These features consist mostly of vehicle and insurance data for customers in 5 US states (the complete data dictionary can be found in Appendix A). This data was clean and required no removal of nulls but did require that the non-numeric columns be converted to numeric for evaluation and modeling purposes.

Table 1: Features

🟥 Customer	🟥 Country	🟥 State Code	⬜ State	🟦 (current) Claim Amount	🟥 Response	⬜ Coverage	⬜ Education	⬜ Effective to Date
⬜ Employment Status	⬜ Gender	🟦 Income	⬜ Location Code	⬜ Marital Status	🟩 Monthly Premium Auto	🟦 Months Since Last Claim	🟦 Months Since Policy Inception	🟦 Number of Open Complaints
🟦 Number of Policies	🟥 Policy Type	⬜ Policy	⬜ Claim Reason	⬜ Sales Channel	🟦 Total Claim Amount	⬜ Vehicle Class	⬜ Vehicle Size	

Table 2: Legend

🟥 Removed Features
🟩 Target
⬜ Kept Features Categorical
🟦 Kept Features Numeric

EDA and Feature Engineering

There are several features that I deemed to be unnecessary in reaching my goals mentioned above. Those features include the customer, country, state code, response, and policy type.

1. The customer feature represents a unique identifier for each row but is not required because there the unique row indexes to do that.
2. The country feature is not necessary because there is only one country in this dataset.
3. The state code is very similar to the state feature.
4. The definition of the response feature could not be determined.
5. The policy type feature is similar to the policy feature.

The numeric features seen above and the target were binned on the median of that particular column. This resulted in a relatively equal distribution of classes in each of the numeric features.

The categorical features were one-hot encoded. These features were one-hot encoded because each subcategory within the categorical columns needed to be scored to determine its overall importance. After one hot encoding and dropping features, the dataset had 61 features.

Answering Goal 1

Determine the most important features in this data set that predicted whether or not the customer would have high insurance.

The determination of the important features was done using KBest. The resulting 15 features after the KBest scoring was

Table 3: Important Features

Basic	SUV	Four-Door Car	Insurance type-Extended	Claim Amount	Insurance type-Premium	Sports Car	Two Door Car
Luxury SUV	Luxury Car	Type of Claim-Collision	Type of Claim-Hail	Type of Claim-Other	Location Type - Suburban		

Answering Goal 2

To use machine learning to determine the demographic of people most likely to have high premiums.

Four ML models were completed in an attempt to answer this question. Since this is a classification problem, classification models were used in answering this question. Finally, the model evaluation was based on the test scores. The results for the Train and Test Scores for these four models are shown below.

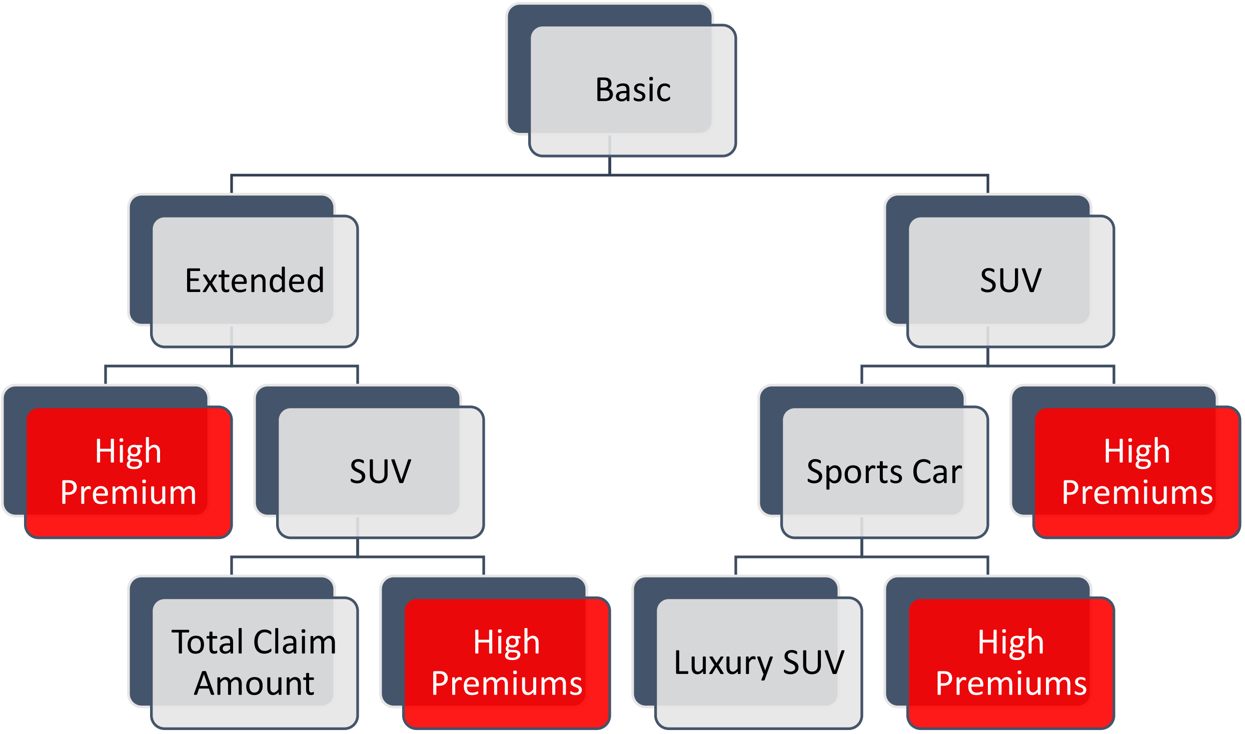
Table 4: Model Scores

Model	Train Score %	Test Score %
Logistic Regression	94.2	94.0
K-Nearest Neighbors	93.9	93.8
Decision Tree	94.1	94.0
Support Vector Classification	94.2	94.0

Results and Conclusions

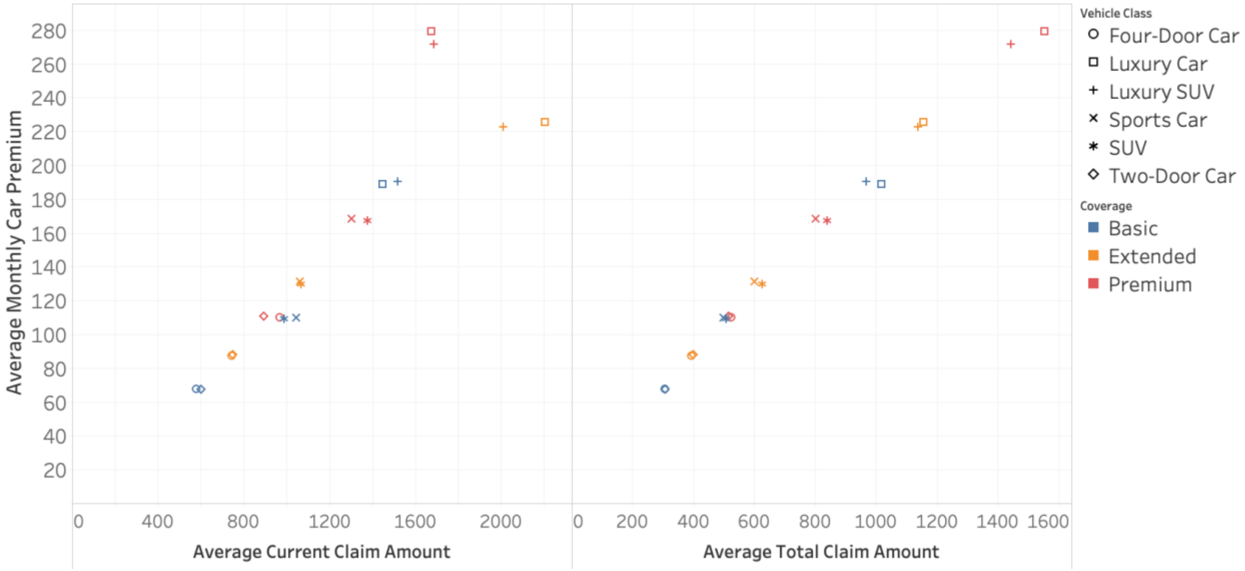
A representation of the important features can be done using a decision tree. The most important features among the 15 found above are the insurance type – basic and extended, SUV, sports car, luxury SUV, luxury car, total claim amount, and location type – suburban.

Figure 1: Sample of the Decision Tree



Interestingly, there is no individual with basic insurance and either a four-door, two-door car with a monthly insurance premium above the threshold with this data set. This was regardless of the total claim amount or the current claim amount, as seen in the visual below.

Figure 2: Comparison of Average Monthly Car Premium to Current and Total Claim Amount



Results of The Decision Tree Model
Accuracy: 94%
Precision: 89%
Recall: 100%

	Predicted Low Premiums	Predicted High Premiums
True Low Premiums	817	110
True High Premium	0	900

From an insurance perspective, the importance of this model is that the model correctly assigned all individuals who are meant to have high premiums. In other words, with this data set, the model was able to predict to 100% accuracy the prediction of high insurance premium individuals.

This model's possible implications to the insurance industry are that, with more customer information data, insurance companies would quickly classify customers/potential customers as high "risk" or low "risk". Classifying potential customers with a model combined with the review of actuaries could potentially result in faster, more accurate approval of insurance services. In addition, more models could be developed to narrow the region in which potential customers could fall into on the scale of the monthly premium.

The next step for this project is to increase the complexity of this classification problem by making it multiclassification instead of binary. Second, it would be to find a more comprehensive feature set with increased customer information.