PURDUE UNIVERSITY | Krannert School of Management

Developing an Interpretable Machine Learning Model in Insurance

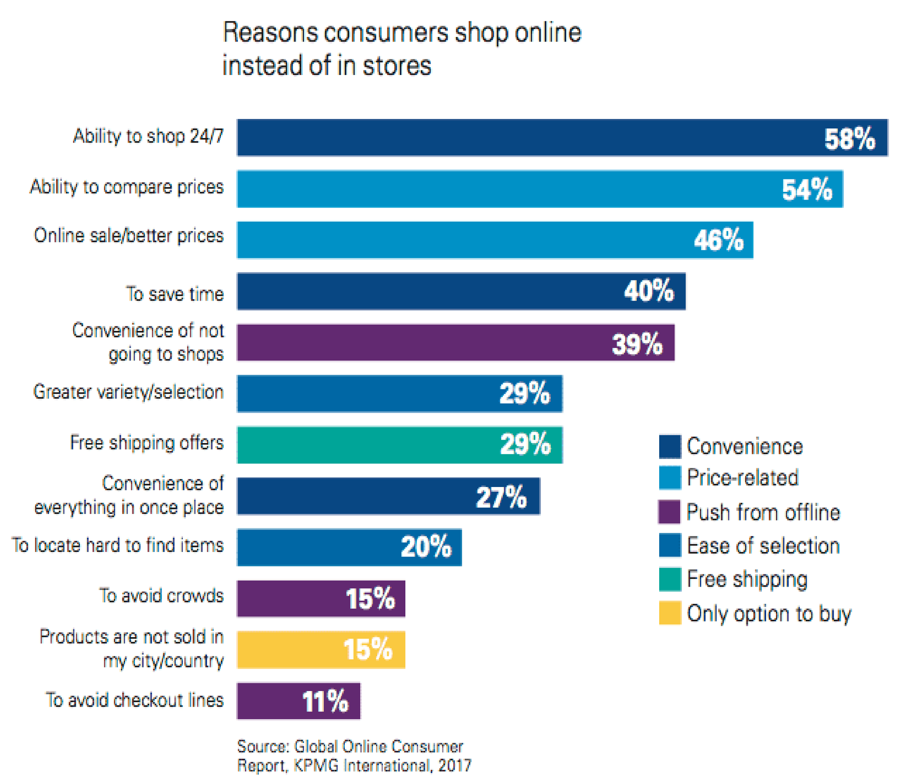
Gurveen Rekhi, Shravaan Bandaru, Jason Denman, Matthew A. Lanham
Purdue University, Krannert School of Management
grekhi@purdue.edu; srbandar@purdue.edu; jdenman@purdue.edu; lanhamm@purdue.edu

ABSTRACT

Our research focuses on developing an accurate and interpretable predictive model in insurance to provide interpretability and transparency to policyholders. The motivation of this study is that the internet has significantly changed how insurance is purchased and pursued. We want to identify how controllable factors such as make, model, and age of vehicle affect the insurance premium. We aim to allow our consumers to understand what goes into calculating the price of these premiums and how they differ for each person uniquely. In order to accomplish our goal, we compile information about a consumer to build a profile and use the profile to build out a unique policy. We posit our approach will provide both premium interpretability and transparency.

INTRODUCTION

The insurance industry accounts for almost 3% of the United States GDP so continuing to grow the industry with more policies is in the best interest of the nation. As consumers have become more conscientious on what they spend their money on it is imperative to make insurance policies understandable to help consumers justify purchasing premiums. By creating a machine learning model that can generate an interpretable model that they can present and explain to a client, we will effectively be making a useful tool that can help increase the amount of premiums sold and increase consumer confidence in their purchase.



https://www.smartinsights.com/ecommerce/ecommerce-strategy/the-reasons-why-consumers-shop-online-instead-of-in-stores/

RESEARCH QUESTIONS

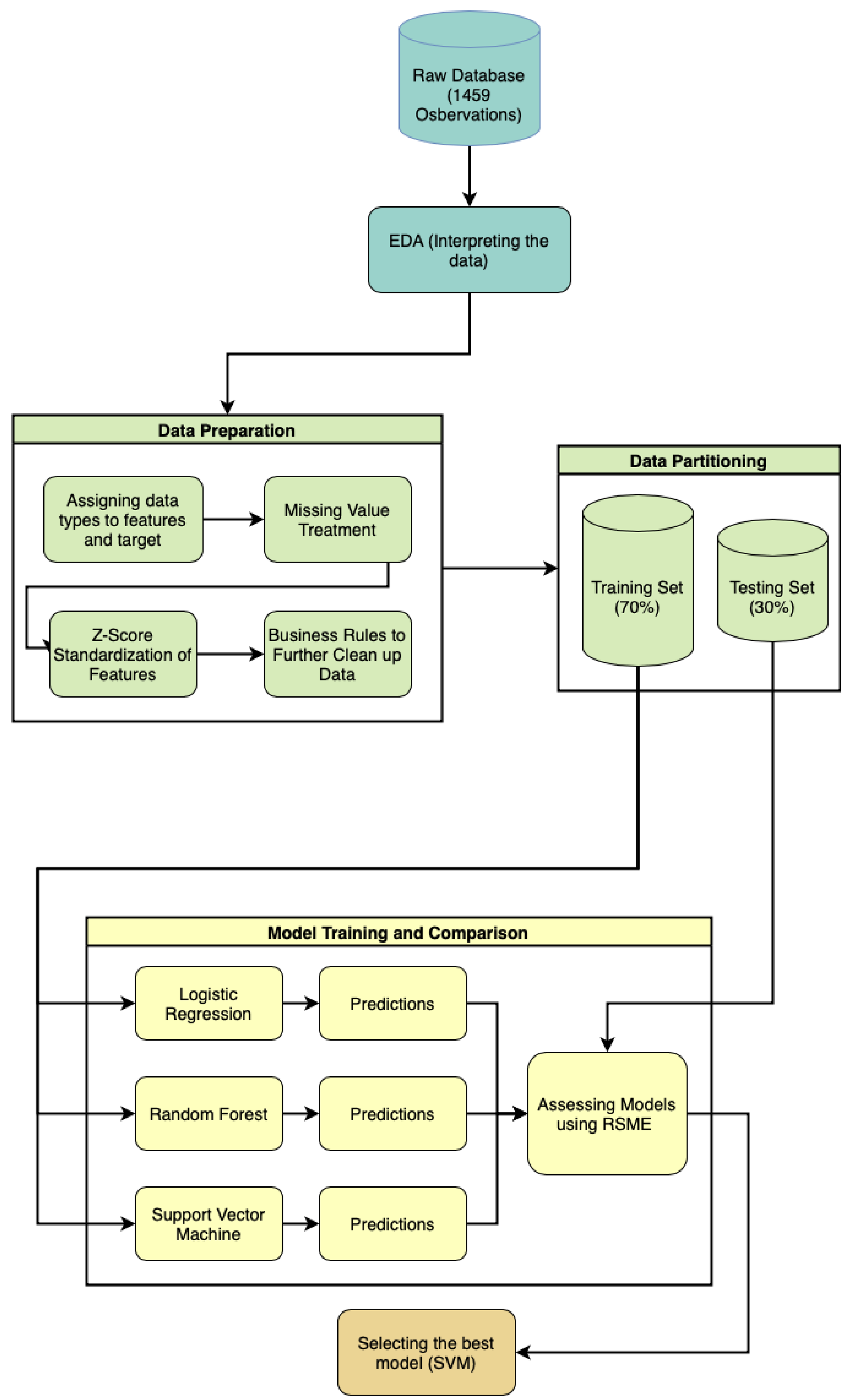
- **Can an interpretable linear regression model be developed for policy holders?**
- **How can we use machine learning techniques (i.e. random forest, SVM) to identify key factors that accurately produce predictions for policyholders and insurance agents?**

LITERATURE REVIEW

We compared papers that explained other factors that are not statistical measures, with our results.

Study	Data that jeopardizes interest	Accessible information	Political Implications	Factors that create Biases
Fung, 2013	X	X		
Hosseini, 2018		X	X	
Schnell, 2017			X	
Barth, 2013			X	
Chen, 2009		X		X
Our Study	X			X

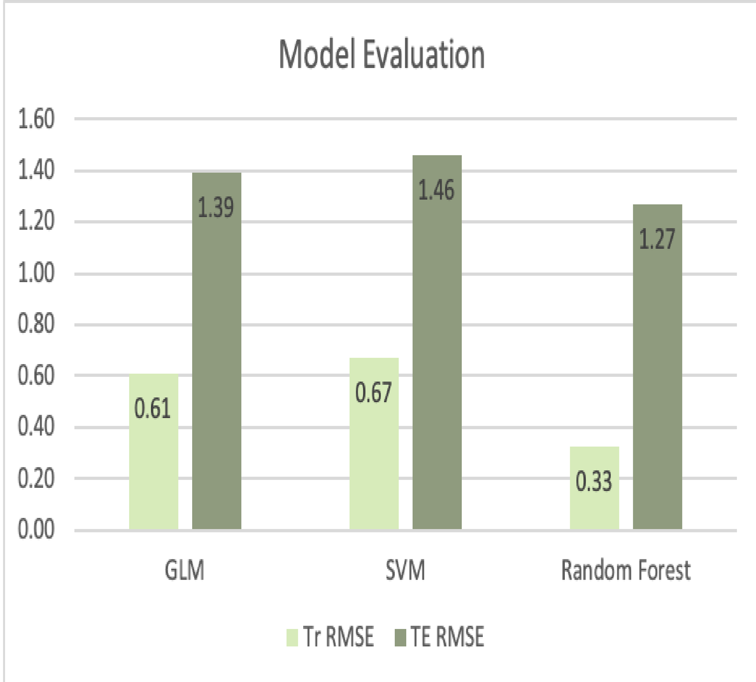
METHODOLOGY

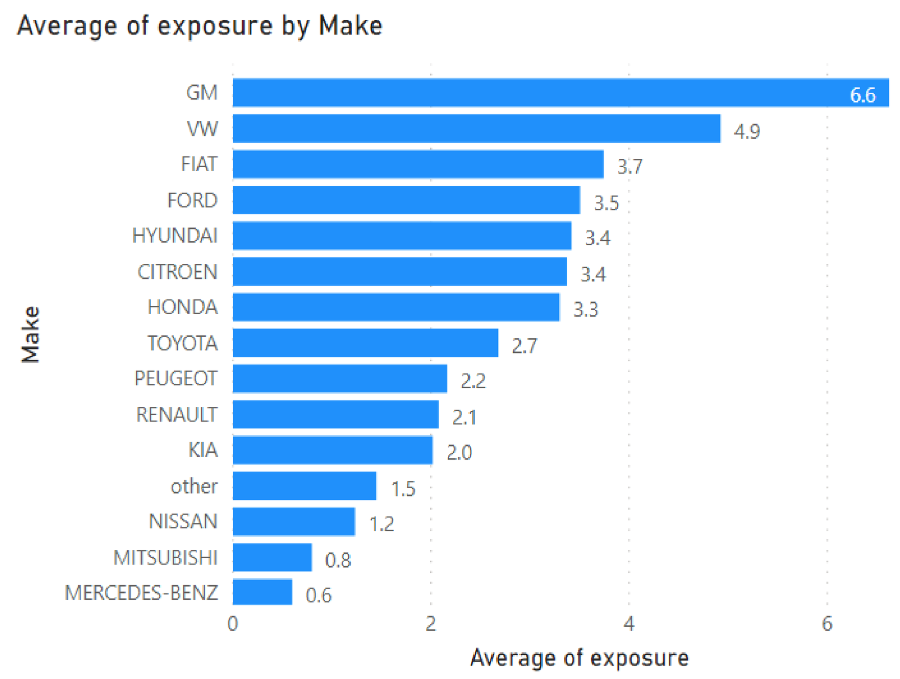


Study Design Figure

Insights

We started off with 6 variables and wanted to see which factors influenced the exposure level. From model training and comparison, we used RSME to assess our models viability and found that linear regression was the best model to answer our research question. Our other research question was to figure out which factors affect exposure. During our data analysis stage we noticed that the brand of the car had the largest impact on exposure (our target variable). Furthermore, Mercedes Benz had the least exposure but upon further analysis we realized that there were very few Mercedes Benz sold. The other interesting insight were the category of vehicles, individuals with pickup trucks, both domestic and import, were deemed safer than sedan drivers. By being male, your exposure rates are higher because in South America there are more male drivers than female.





Business Impact: Based on the profile that is built around a customer, we can use past data to compare similar profiles using the contributing factors like sex, age, and vehicle type. Using a linear regression, we can show an easily interpreted model to a client and explain the relationship between their profile and historic data that explains the drivers behind their premiums cost.

This can be a useful tool for insurance agents to sell more affordable policies by:

- Showing the drivers behind the cost of the policy so the customer can understand how to lower the cost
- Determine if insuring a customer is a good choice based on their profile
- See the profiles of customers that are low-risk and offer them incentives to join or switch to your insurance agency

CONCLUSIONS

We found that having a restricted number of predicting variables and the dataset being from a foreign country was a limitation to accurately predict our target variable. The main variables that influenced the results are the type of vehicle customers owned and gender. Going forward, having a local data set that originated from the United States with more predictive variables such as average repair costs would be advantageous in increasing the accuracy of the results.

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham and our graduate student mentors student Xinyu Wang and student Theo Ginting for their guidance and support on this project.