Predicting Claim Amounts in Auto Insurance

04/30/2025

STOR 320

Group 6

Introduction

- Auto insurance companies offer financial protection to mitigate risks in exchange for a premium payment.

- Data analytics on risk management
  - Predict potential risk factors
  - Access client profiles
  - Refine pricing models

- **Our project:**
  - Using data on car insurance policyholders and accidents to make predictions about auto insurance claims

Question:

Can we predict the total claim amount using policyholder tenure, deductible amount, policyholder age, and the time the incident occured?

Datasets

- **Kaggle's Auto Insurance Claims Dataset**
  - 1000 records of insurance claims from 2015
  - 39 usable features:
    - months as a customer, policy deductible, incident state, injury claim, collision type, etc.

- **U.S. Federal Highway Administration State Motor-Vehicle Registrations 2015 Dataset**
  - 50 state-level rows, 15 features of registration counts
  - Merged with claim data via **state**

Question:

Can we predict the total claim amount using policyholder tenure, deductible amount, policyholder age, and time of day?

Variable	Description	Type	Example
total_claim_amount	Final amount in USD paid out by the insurance company	Numerical	$52.8k
months_as_customer	How long the policyholder has had coverage	Numerical	204
policy_deductible	Deductible amount chosen by policyholder in USD	Categorical	[$500, $1000, $2000]
age	Age of policyholder	Numerical	38
incident_hour_of _the_day	Hour when the incident occurred	Numerical	11
registration_category	Total vehicle registration category by state: top 25% labeled "High," others "Low.”	Categorical	“High” or “Low”

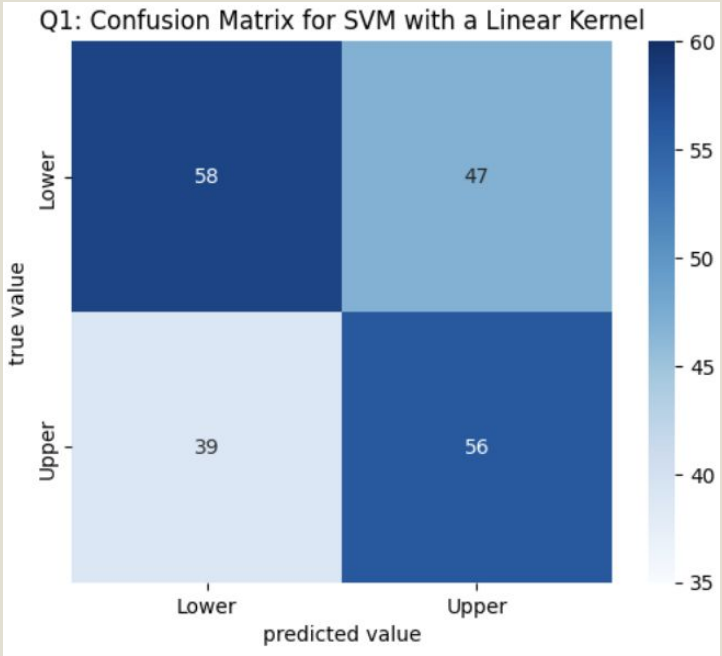
Modeling Approach

- Regression → Classification
- Binary classification task: High vs. Low total claim amounts
- Baseline model that always predicts upper claims
  - Accuracy: 47.5%
- Models tested:
  - *Linear Regression*
  - *Random Forest Regressor and Classifier*
  - *Logistic Regression*
  - *Naive Bayes*
  - *Decision Tree*
  - *Support Vector Machines*
- Train/test split: 80/20
- Evaluation metrics: Accuracy, Precision, Recall, F1-Score

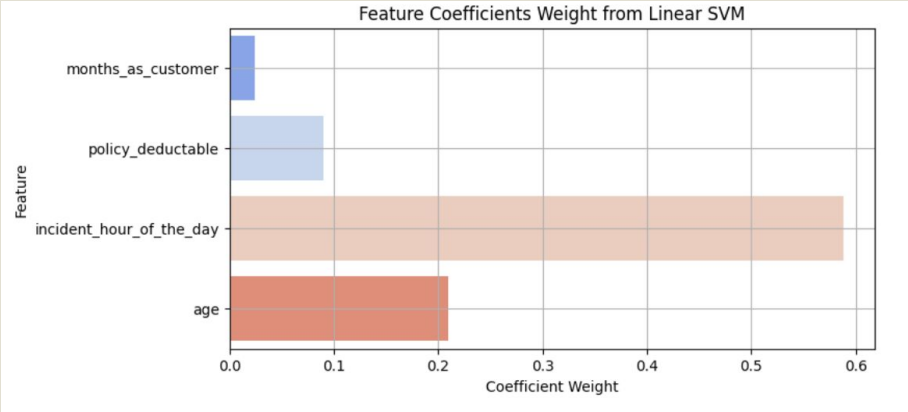
Best Model

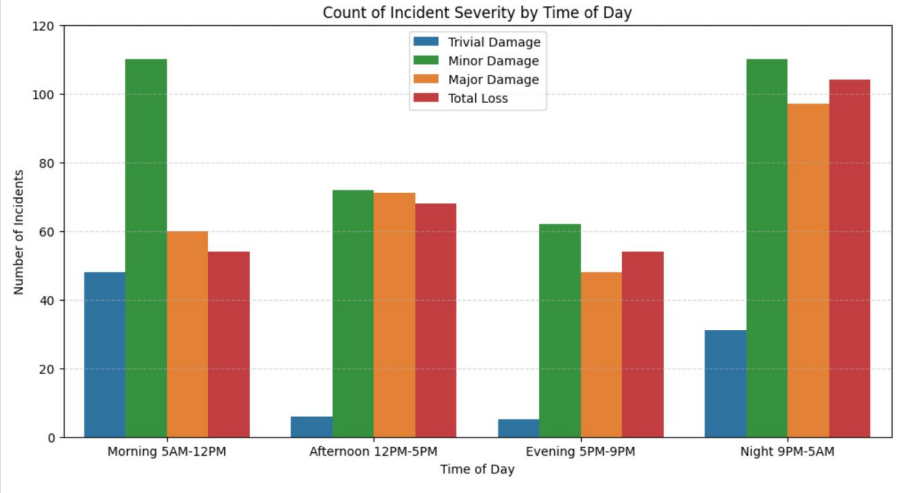
- Support Vector Machine with Linear Kernel
  - Gridsearch tuning hyperparameter C = 1000
  - Classification label:
    - 0: lower 50% of total claims
    - 1: higher 50% of total claims

	Class	Precision	Recall	F1-Score	Overall Accuracy
0	Lower	0.60	0.55	0.57	0.57
1	Upper	0.54	0.59	0.57	



Interpretation of Model





Practical Insights and Conclusion

- Little evidence to suggest policyholder tenure, policy deductible, the time of day the incident occurred, and age are good predictors of total claim amount.
- Only relevant to the 3 states the policyholders are from and the 7 incident states in our dataset
- Future work:
  - Backwards parameter selection?
  - Using more external factors?
  - Different models?

Works Cited

Bondaug-Winn Nick, “20 Key Risk Mitigation Techniques for Auto Insurance Agencies”. https://www.hbwleads.com/blog/key-risk-mitigation-techniques-for-auto-insurance-agencies/

North Carolina Department of Insurance. “Auto and Vehicle Insurance”. https://www.ncdoi.gov/consumers/auto-and-vehicle-insurance

Shah, B. (2018, August 20). *Auto insurance claims data*. Kaggle. https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data

Table MV-1 - highway statistics 2015 - policy: Federal Highway Administration. Table MV-1 - Highway Statistics 2015 - Policy | Federal Highway Administration. (n.d.). https://www.fhwa.dot.gov/policyinformation/statistics/2015/mv1.cfm

Thank you!