STOR 320: Introduction to Data Science

Spring 2025

Final Paper Group 6

Introduction

Insurance companies rely on data to guide pricing, risk assessment, and claims processing decisions. As access to detailed customer and incident data expands, so does the opportunity to use data science not just to explain past outcomes, but to make informed predictions about future ones. This project explores how data features related to car insurance policyholders and car accidents can be used to predict important aspects of auto insurance claims, such as how costly a claim might be or whether it will result in a total loss, indicating a totaled vehicle. These kinds of predictive insights can help insurers operate more efficiently, personalize services, and improve transparency for customers. Similar research has been explored in international contexts as well. For example, a study on car insurance claims in Athens, Greece, incorporated external variables such as weather conditions and car sales to enhance predictive power (Poufinas et al., 2023). This inspired us to consider broader contextual factors, like vehicle registration by state, in our own analysis.

To conduct our analysis, we combined two related datasets. The first dataset contained detailed information from the United States about individual insurance claims from 2015, including variables like claim amount, deductible, incident severity, policyholder age, tenure with the company, and the time of day the incident occurred. The second dataset offered state-level data from the United States on vehicle registration of publicly, commercially, and privately owned vehicle types from 2015, giving us a way to account for broader geographic patterns in vehicle usage and traffic exposure. Together, these sources allowed us to examine both personal and environmental factors that could influence insurance claim outcomes. We do note that our data comes from 2015, which we recognize is not recent and presents a limitation in terms of capturing and generalizing our results to current insurance trends and behaviors.

As our modeling progressed, we noticed that our original research questions needed adjustment. Some models, especially those attempting to predict the exact claim amount, performed poorly due to weak correlations and high variance in the target. In response, we refined our questions to focus on more manageable, classification-based outcomes. Our updated research questions became: (1) Can we predict whether a total claim amount will fall in the high or low range using policyholder tenure, policy deductible, the time of day the incident occurred, and age? and (2) Can we predict whether an incident will result in a total loss using the time of day, policyholder age, tenure, and deductible? These questions allowed us to shift from simply identifying trends to testing whether those trends could drive meaningful

predictions. They also framed our analysis to align with real-world applications in the insurance industry, where classifying claims by severity or risk is often more useful than estimating exact dollar amounts.

Through these investigations, we aimed to demonstrate how data science techniques can uncover valuable insights from routine insurance data. Our goal was not only to build models with strong performance but also to understand what those models reveal about customer behavior, risk, and the factors that drive severe or costly claims. For the insurance companies that collect and use this data, these insights can help them spot high-risk claims sooner, speed up the claims process, and make better decisions about how they price their insurance plans. We hope our work can contribute to insurance companies making data-driven decisions and providing a better experience for policyholders.

Data

Our primary dataset comes from Kaggle, though the original creator is not clearly identified. An individual with the name Bunty Shah uploaded the dataset. However, Shah noted that he was not the original creator of the dataset and had added the dataset from an external source. The dataset contains United States auto insurance claims from the year 2015, with a majority of the policyholders in Ohio, Indiana, and Illinois, and incidents that occurred in several other states, including South Carolina, Virginia, New York, Ohio, West Virginia, North Carolina, and Pennsylvania. The sample includes 39 usable variables and 1000 observations, each representing a single insurance claim made by a policyholder. The population would be all auto insurance claims filed in 2015 in the U.S. states represented in the dataset, primarily Ohio, Indiana, Illinois, and to a lesser extent, South Carolina, Virginia, New York, West Virginia, North Carolina, and Pennsylvania.

This dataset includes a mix of categorical and numerical variables related to the customer, their insurance policy, the accident, and the claim itself. Key variables used in our analysis include: *age* (age of the policyholder), *months_as_customer* (tenure with the insurance company), *policy_deductible* (the chosen deductible by the customer–either 500, 1000, 2000 US Dollars. Corresponds to what the customer will pay up front when an accident happens), *incident_hour_of_the_day* (hour the incident occurred), *incident_severity* (severity of the incident categorized as trivial damage, minor damage, major damage, and total loss), *total_claim_amount* (total claim filed for accident in $). One limitation to note is that the *incident_severity* variable is somewhat vague, especially for categories other than "Total Loss." While "Total Loss" clearly indicates a totaled vehicle, the definitions of other severity levels are less specific and may vary in interpretation. The data did not have any missing values that we had to account for, however, there was one column named c_39, which only had missing values, which we omitted from our data analysis due to its lack of information on its relevance. The data was very well cleaned and filtered before our analysis, so we did not have to do additional cleaning and filtering.

To supplement our analysis, we used a second dataset from the Federal Highway Administration, which provides United States state-level motor vehicle registration statistics from 2015. To combine both datasets, we performed an inner merge between the insurance claims dataset and the Federal Highway Administration dataset using the state abbreviation columns. This dataset includes the total number of registered vehicles in each state, broken down by vehicle type and ownership. Vehicle types included

automobiles, buses, trucks, motorcycles, and all motor vehicles. Ownership included private and commercial, publicly owned, and total. We chose this source specifically because it aligns with the year of our claims dataset and considers broader contextual factors. We added a variable to our dataset named *registration_category* to classify states based on vehicle density to further our analysis. States with total vehicle counts above the 75th percentile were labeled as "High," while the rest were labeled as "Low." The only state from our dataset that was classified as “High” was New York, which is a limitation in generalizing our findings to “High” density states.

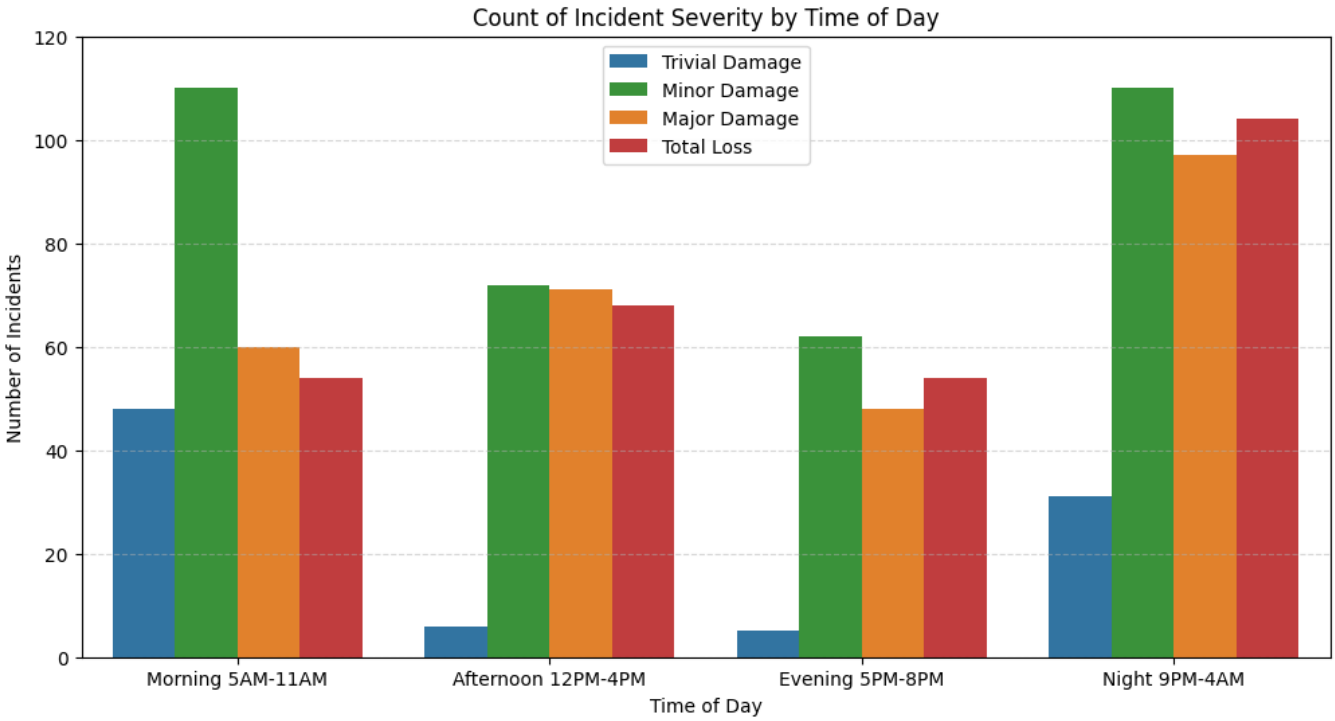
The data had no missing values; however, some of it was estimated due to incomplete state reporting. Even so, the dataset still added valuable context to our analysis. Our analysis focused particularly on the number of publicly owned vehicles and the total number of registered automobiles in each state, which matched the states present in our primary dataset. This allowed us to initially ask questions like whether states with more publicly owned vehicles experienced different patterns in claim frequency or severity, which we explored in our exploratory data analysis using the *registration_category* variable we created. It allowed us to examine how state-level factors, such as the number of publicly owned vehicles or total registered automobiles, might be related to claim severity or the timing of incidents. Because it aligned with the year of our primary dataset, it was a strong fit for drawing meaningful comparisons across both individual and regional levels.

The descriptive table includes all of the variables of interest from our merged data from Kaggle and the Federal Highway Administration. It provides a short description of each variable to clarify understanding. It also includes the possible values or ranges for each variable. The categorical variables display the possible values or categories the variable contains. The numerical variables display the range of values, meaning the minimum and maximum. The last column displays the units for numerical variables and any final notes for the categorical variables.

The descriptive figure we chose was a bar chart showing the number of incidents by time of day and incident severity. We were interested in seeing if an external variable like the time of day had any relationship with incident severity. To make the chart more readable, we created a new variable called *time_of_day* by grouping the hour of the incident into broader time categories: Morning (5 AM–11 AM), Afternoon (12 PM–4 PM), Evening (5 PM–8 PM), and Night (9 PM–4 AM). The *time_of_day* variable was only used in this visualization for readability, and *incident_hour_of_the_day* was used instead for the rest of our analysis. We found that Minor Damage is the most frequent severity across all time periods, especially in the morning. Total Loss incidents occur most often at night. Trivial Damage is relatively rare across all time slots. Overall, nighttime sees the highest number of incidents, which we noted in the creation of our research questions and model creation.

Description of Selected Variables

	Variable	Description	Possible Values/ Range	Units/Notes
0	months_as_customer	Number of months the customer has held a policy	0 to 479	Months
1	age	Age of the customer	19 to 64	Years
2	policy_deductable	Deductible amount the policyholder choose	500, 1000, 2000	U.S. Dollars
3	registration_category	Vehicle registration category by state	Low, High	Based on total vehicle registration volume
4	total_claim_amount	Total amount claimed for the incident	100 to 114,920	U.S. Dollars
5	incident_hour_of_the_day	Hour when the incident occurred	0 to 23	24-hour clock format
6	incident_severity	Severity of the reported incident	Trivial, Minor, & Major Damage, Total Loss	Categorical description of damage level



Results

For reference, our first research question was: Can we predict the total claim amount using policyholder tenure, deductible amount, policyholder age, and the hour of the day the incident happened? In the data section, we mentioned an encoded categorical version of *hour_of_the_day.* We tested this variable in place of *incident_hour_of_the_day,* but it performed worse, so we ultimately decided to use the original numeric version. We did not perform any additional feature selection beyond our initial EDA, as the

variables used *months_as_customer, policy_deductible, age,* and *incident_hour_of_the_day* were identified early on as the most relevant predictors based on our exploratory analysis.

We began with a baseline model using the mean of our target variable, which, as expected, did not perform well, showing a low R-squared value of -0.010 and a high root mean squared error (RMSE) of 25,928.38. We then attempted to fit a linear regression model, which performed only slightly better than the baseline, with a very low R-squared value of 0.0512 and a high RMSE of 25,128.60. This outcome was expected, as our exploratory analysis suggested that the assumptions of linearity were not met. Although we didn't observe clear signs of heteroscedasticity, the correlation matrix showed very weak relationships between variables, and the data did not appear to follow a normal distribution. With this, we then pivoted to a Random Forest Regressor, but the results were actually worse, yielding an even lower R-squared and higher RMSE than the linear regression model. Since both regression models performed so poorly, we made a larger pivot and reframed the problem as a classification task instead.

To do this, we converted the *total_claim_amount* variable into three categories: lower, middle, and upper thirds. This transformation resulted in a naturally balanced dataset, which is ideal for classification models. We first tried a Random Forest Classifier and achieved an accuracy of 43.5%, but precision and recall scores for all three classes were under 47%. We then tested a Support Vector Machine (SVM) with a linear kernel, but the accuracy dropped slightly to 42.5%, and precision/recall remained similar. We also tried a radial basis function (RBF) kernel, but this yielded slightly worse results.

We suspected that having three classes was making the classification task more difficult, so we simplified the target to just two categories: lower 50% and upper 50%. We trained a baseline model that always predicts upper claims, since these are the more important cases in our analysis. Although the dataset was perfectly balanced, the baseline model achieved an accuracy score of 47.5%, due to a slight imbalance in the test split. When we retrained a Random Forest Classifier on this version, the accuracy improved to 54.5%, along with better precision and recall scores. We then tried an SVM with a linear kernel and achieved our best results: 57% accuracy and stronger precision and recall across both classes. Along with F1-scores of 57% for both classes. We also tested an SVM with an RBF kernel, but the performance dropped slightly. We used GridSearchCV for hyperparameter tuning of the SVM model and found that a linear kernel with a C value of 1000 yielded the best results. We also tried Logistic Regression and Naive Bayes, but neither model outperformed the SVM with the linear kernel. While the Naive Bayes model had a slightly higher F1-score of 61% for class 0, its score for class 1 was only 51% and class 1 is more important in this context, as insurance companies are especially interested in identifying higher claim amounts, so we chose to prioritize class 1 in our model choice.

Based on these results, we selected the Support Vector Machine with a linear kernel and C=1000 as the final model for this question. It achieved the highest overall accuracy at 57% and the most balanced F1-scores, with 57% for both classes. We included a table displaying these numbers, also known as the classification report. Additionally, we included a confusion matrix to illustrate how well the final model predicted class 0 (lower 50%) and class 1 (upper 50%), highlighting both correct classifications and common misclassifications. While our model didn't achieve high performance metrics, we explored multiple approaches and made the most of the available data. Limitations included a small dataset, potential subjectivity in target variable definitions, old data, data from only one year, and data from only

select states.

For our second research question, we asked: Can we predict whether an incident results in a total loss using policyholder age, deductible amount, witnesses, and the hour of the day the incident happened? Total loss claims are often more expensive and time-sensitive, as we saw in our exploratory analysis, so being able to predict them early could help insurers allocate resources more efficiently and improve the claims process. Although the question originally included the number of witnesses as a predictor, we removed it during feature selection. Our exploratory analysis showed that including the witnesses variable consistently reduced model accuracy, precision, and recall across all classifiers. As a result, we proceeded with four stronger features: *incident_hour_of_day, age, months_as_customer,* and *policy_deductible*.

To approach this, we created a new binary target variable called *is_total_loss*, where a value of 1 indicated a total loss and 0 otherwise. The *is_total_loss* variable was heavily imbalanced to try to counteract this, we added class_weight='balanced' in all of our models. We trained a baseline model that always predicts total loss claims, since these are the more important cases in our analysis. Given the highly imbalanced dataset, with total loss cases making up only 28% of the data, this baseline achieved an accuracy score of 22%. We then trained a Random Forest Classifier using our selected features. The model achieved a relatively high accuracy of 75%, but only 23% recall for class 1 (total loss), indicating that the model mostly predicted non-total loss cases and struggled to correctly identify actual total loss claims. We potentially saw this as overfitting. We then added a feature called *registration_category,* which is described in the first table. However, this addiction did not improve performance; the recall for total loss fell to 20%, and accuracy slightly decreased to 74%.

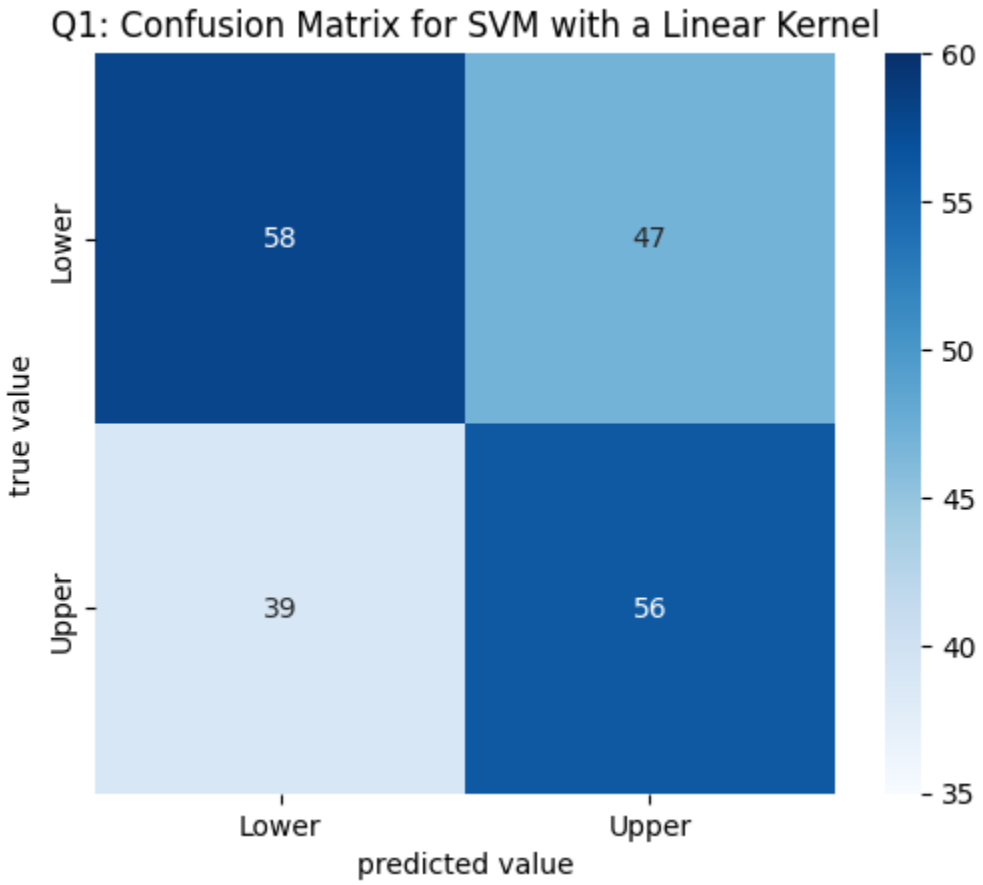
Given these results, we shifted to an SVM with an RBF kernel, which often performed better on smaller or imbalanced datasets. Although the overall accuracy dropped to 56.5%, the model achieved a much stronger recall of 66% for total loss claims, making it far more effective for our specific goal. Since total loss claims were the minority in our dataset, recall was prioritized over accuracy as our main evaluation metric. We included a table displaying the model's performance, also known as the classification report, where the precision and recall scores can be found. Additionally, we included a confusion matrix to illustrate how well the model predicted class 0 (not total loss) and class 1 (total loss), highlighting both correct classifications and common misclassifications. To further optimize performance, we used GridSearchCV to tune hyperparameters and found that the best performance came from an RBF kernel with C=5 and gamma = 'scale'. This combination allowed the model to maintain balance across both classes while identifying significantly more of the true total loss outcomes.

In addition to SVM, we also tested a Logistic Regression model with and without the *registration_category* variable. Logistic regression offered the benefit of interpretability and a relatively fast runtime, but in our case, its performance was limited. Both versions produced an accuracy of 50.5%, with a recall for class 1 around 48%. While this was better than the Random Forest's recall, it still did not outperform the SVM model. These results suggest that logistic regression may not have been flexible enough to capture the more complex interactions in our features, especially the nonlinear influence of variables like incident time and tenure.

To complete our analysis, we evaluated a Decision Tree Classifier using the same four main features. This model produced an accuracy of 51.3% and a recall of 51% for total loss cases. While this performed slightly better than logistic regression in terms of identifying class 1 instances, it did not match the SVM's performance because of its relatively lower recall. After evaluating all models, we concluded that the Support Vector Machine with an RBF kernel offered the strongest performance overall. It consistently delivered the highest recall for the minority class while maintaining a reasonable level of balance across both classes. In the context of real-world claims where catching costly, severe incidents is a top priority, this tradeoff made the SVM the most effective model for answering our research question. While our results are not great, we did our best with the data available and the imbalance in the *is_total_loss* variable.

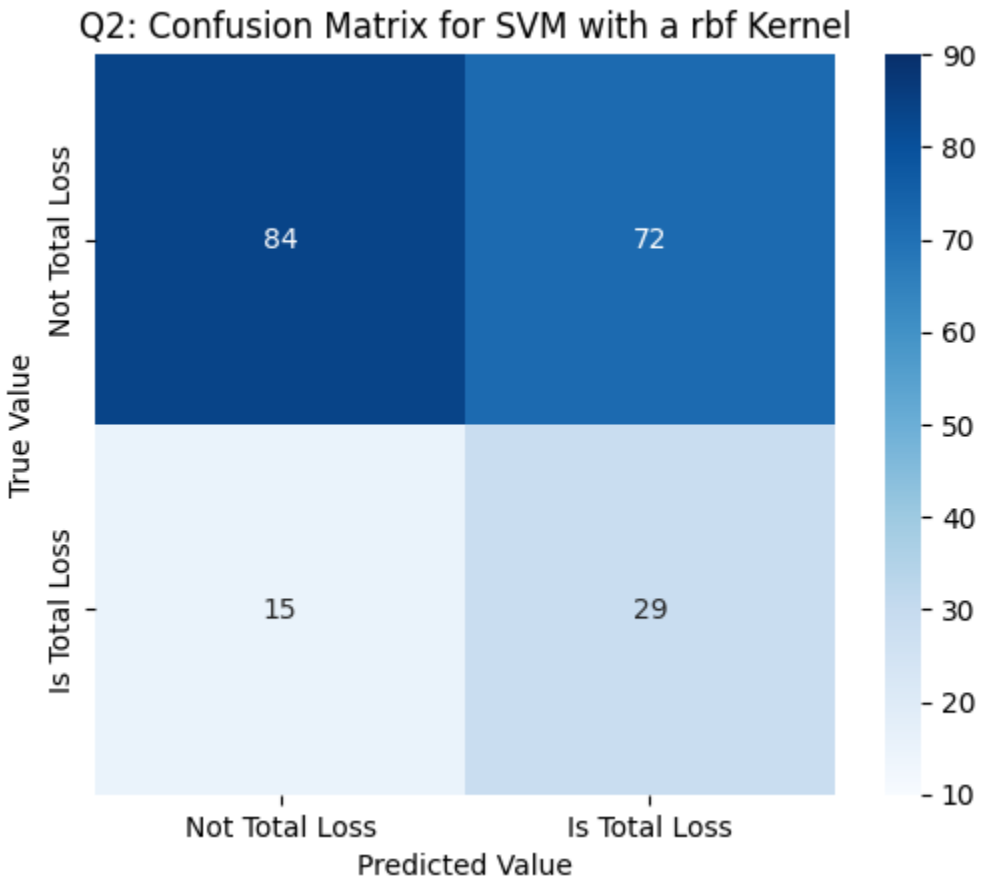
Partial Classification Report for Q1

	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	



Partial Classification Report for Q2

	Class	Precision	Recall	F1-Score	Overall Accuracy
0	Not Total Loss	0.85	0.54	0.66	0.565
1	Is Total Loss	0.29	0.66	0.40	



Conclusion

At the start of our project, we aimed to answer two key questions: Can we predict whether a claim amount will be high (upper 50%) or low (lower 50%) using features like policyholder tenure, deductible, policyholder age, and time of day? And can we predict whether an incident will result in a total loss using the variables policyholder age, deductible amount, witnesses, and the hour of the day of the incident? These questions came from patterns we noticed during our initial data exploration and were designed to reflect real-world challenges that insurance companies face. While we began by trying to predict exact claim amounts with regression models, we quickly found that the results were weak, with low R-squared values and high error rates. So, we shifted to classification models, which performed better. For the first question, a Support Vector Machine with a linear kernel gave us the best results with 57% accuracy. For the second, an SVM with an RBF kernel gave us a 66% recall on predicting total loss, which was a big

improvement over the other models. While it is possible to make predictions for both outcomes, the performance metrics indicate that these models may not be the most reliable, and alternative modeling approaches would be needed for real-world use.

Even though our models didn't hit high marks in terms of accuracy, they gave us useful insight into what factors mattered most. Time of day and deductible consistently showed up as relevant predictors, while features like witnesses didn't help and even hurt performance in some cases. That finding itself is important, knowing what not to include can be just as valuable. In the real world, insurance companies care more about catching serious claims that result in high amounts and totaled vehicles early rather than predicting exact dollar amounts, which is why we focused on recall for total loss predictions. While we didn't build a perfect model, our findings suggest that even simple customer and incident data can offer meaningful signals for identifying high-risk claims, which could help insurance companies prioritize resources more effectively.

Looking ahead, there's a lot of room to grow in this work. Having access to more recent data or broader geographic coverage would help improve the reliability of our results. It would also be interesting to test models that include behavioral data, like past claim history or driving violations, which we didn't have. On the modeling side, we'd like to try more advanced techniques like those used by Wilson, Nehme, Dhyani, and Mahbub (2024), who compared Generalized Linear Models (GLMs) with Gradient Boosting Machines (GBMs), Artificial Neural Networks (ANNs), and a hybrid GLM-ANN model. These techniques helped them improve prediction accuracy by capturing complex relationships in the data while considering interpretability and performance. We also ran into challenges with class imbalance, especially with total loss cases, and would want to explore better strategies for handling that. Still, this project helped us understand the strengths and limits of predictive modeling in insurance and showed us how even small improvements can make a difference in real-world applications.

References

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

Poufinas, T., Gogas, P., Papadimitriou, T., & Zaganidis, E. (2023). Machine learning in forecasting motor insurance claims. Risks, 11(9), 164. https://doi.org/10.3390/risks11090164

Wilson, A. A., Nehme, A., Dhyani, A., & Mahbub, K. (2024). A comparison of generalised linear modelling with machine learning approaches for predicting loss cost in motor insurance. Risks, 12(4), 62. https://doi.org/10.3390/risks12040062