**Predicting Road Accident Risk**
GB 657 - Final Project Report
Jakob Noll, Miles Esguerra, Mai Nguyen, Matthew Popp

---

**Competition URL:** https://www.kaggle.com/competitions/playground-series-s5e10/overview
**Colab Notebook URL:** GB657 Final Project - Group 1.ipynb

**I. The Business Case:**
In 2025, the National Safety Council (NSC) reported a 13% drop in traffic fatalities, the lowest in years despite increased driving, yet more than 17,000 lives were still lost and nearly $1.8 trillion in societal harm occurred. Even with these improvements, meaningful road safety requires more proactive strategies. Acting on behalf of the state government, we are implementing initiatives to strengthen the Safe System approach, which seeks to eliminate fatalities by improving safety across people, roads, vehicles, speeds, and post-crash care. While the Safe System provides a strong foundation, traditional methods often leave agencies reacting to crashes rather than anticipating them.

To address this, we have developed a predictive model that analyzes environmental, infrastructural, and temporal factors to identify road segments most likely to experience accidents before they occur. This enables us to prioritize interventions based on data-proven risk rather than intuition. Maintenance teams can focus repairs and signage where they will have the greatest impact, transportation departments can schedule patrols or temporary closures proactively, and policymakers gain clearer insight into long-term patterns shaped by weather, traffic flow, and infrastructure age.

By applying predictive insights, our work enhances the Safe System approach, catching risks early and directing resources where they will save the most lives and reduce costs. The model does not replace Safe System principles but sharpens them, providing measurable, data-supported action items. Predictive accuracy is assessed through Root Mean Squared Error (RMSE), giving agencies confidence in the model's reliability. Strong, consistent performance builds trust, supporting sustained partnerships and creating opportunities to expand into related areas such as environmental risk modeling and transportation optimization.

**II. Data Analysis and Preprocessing:**
The dataset used in this project was obtained from Kaggle and was derived using a deep learning model trained on simulated road accidents. The training and tests set were generated from that model, resulting in feature distributions that are similar to those found in the original dataset. The original training portion of the dataset contained over 517K records with 12 initial features.

An examination of the categorical features show the distribution of key environmental and roadway conditions is relatively balanced. Weather types such as clear, rainy, and foggy occur in

similar proportions, ensuring that the model is not overly influenced by any single weather condition. A similar pattern is observed in road types, where urban, rural, and highway segments are well represented, allowing the model to learn accident-risk patterns across diverse roadway environments.These categorical features were one-hot encoded so they could be represented in our model.

Lighting conditions such as daylight, dim, and nighttime also appear in roughly equal frequencies, providing comprehensive coverage of visibility scenarios. This balanced distribution across multiple categorical variables helps reduce bias and supports the development of a more generalizable accident-risk prediction model.

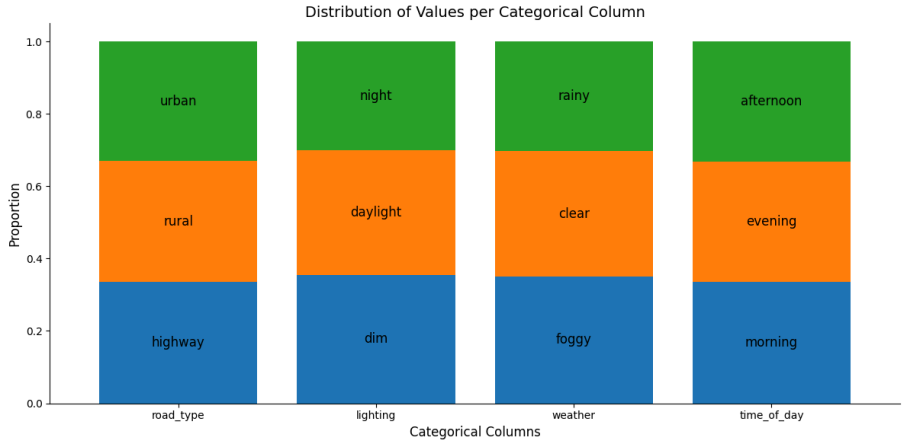



**Figure 1: Distributions of Categorical Features**

The prediction variable of "Accident Risk" follows a relatively normal distribution, as shown in figure 2. This balanced spread contains a healthy mix of low, medium, and high-risk scenarios. This data is helpful in modeling since it's not overly skewed, the model can learn smoother and more generalized patterns.

Feature engineering was implemented to enhance the models ability to capture complex relationships in the data and improve predictive accuracy. Three engineered features were added to capture conditions that increase driving risk, including dangerous curves, adverse weather, and reduced lighting. We applied a scaler to the curve danger variable, as well as the number of lanes and speed limit so that all of the feature variables would either be binary or have a mean of 0 to reduce model bias based on scale.

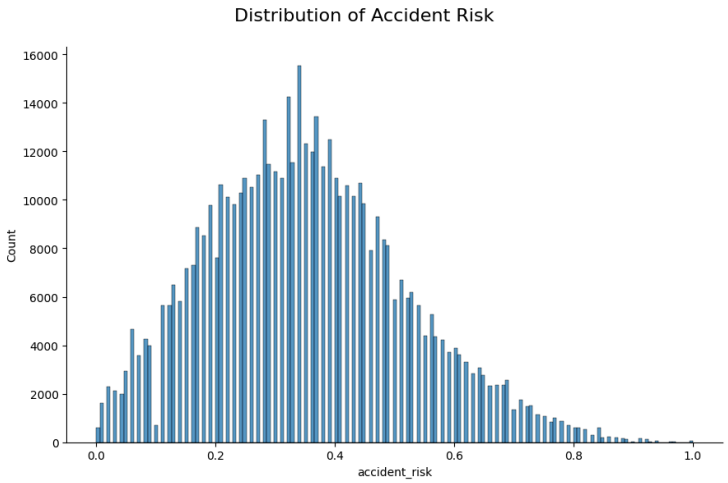



**Figure 2: Distribution of Accident Risk**

## III. Modeling & Results

We built and trained four main models in an effort to find the optimal machine learning model to predict accident risk. In all of our models cross validation was used to prevent overfitting, and fit data based on generalizable patterns. We primarily evaluated models for accuracy and overfitting based on $R^2$ and RSME, which was the metric used to judge the Kaggle competition. The first model ran was an OLS Regression model that we used as a baseline to compare to other models. The baseline model performed worse than the more complex models, with an $R^2$ of 0.8051 and a RMSE of 0.0735.

Next, we trained a Random Forest model with cross validation across 5 folds to catch more complex nonlinear relationships and feature interactions. It was able to achieve 0.07 higher $R^2$ and 0.018 lower RSME than the baseline model. It took about 3.5 minutes to evaluate the model across all of the folds, which was much longer than the 1 second baseline fitting and evaluation.

Following the Random Forest, a Gradient Boosting Regressor was trained to evaluate whether sequential boosting of weak learners could better capture patterns in the data. The gradient boosting model performed better than the Baseline OLS model, but about the same as the untuned Random Forest achieving an $R^2$ of 0.881 and a RMSE of 0.057. Notably, the Gradient Boost model took about 30% longer to train than the Random Forest.

Finally, a neural network regression model was implemented using both Tensorflow Keras and Pytorch, which had similar performance. The architecture consisted of 2 hidden layers using ReLu activation and was trained for 20 epochs to analyze complex patterns in the data. This Neural Network performed nearly identical to the Gradient Boosting Model, with a $R^2$ of 0.883 and a RMSE of 0.057. Interestingly, the deeper neural network models with more hidden layers tended to perform worse than the less complex neural networks, and took significantly longer to train.

We chose to tune our Random Forest model to further enhance its performance. We began by running a randomized search with cross validation over key parameters such as the depth, number of features, and number of estimators. This search returned the optimal parameters of 100 estimators, max depth of 10, minimum sample per split of 5, and minimum samples per leaf of 1. This random search took 45 minutes. We later ran a full grid search with cross validation which yielded the same results, aside from recommending 200 estimators. The grid search took roughly twice as long as the random search, so this is a good example of how random search can be used to achieve similar results to grid search more efficiently.

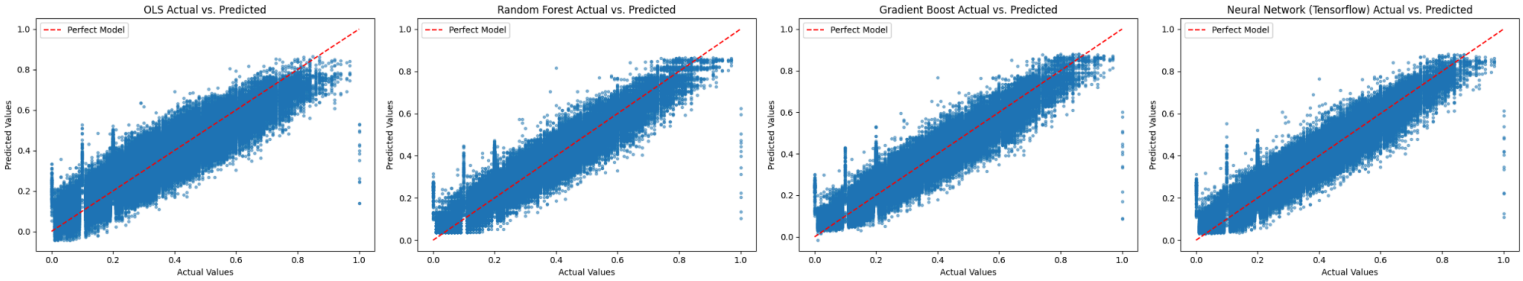
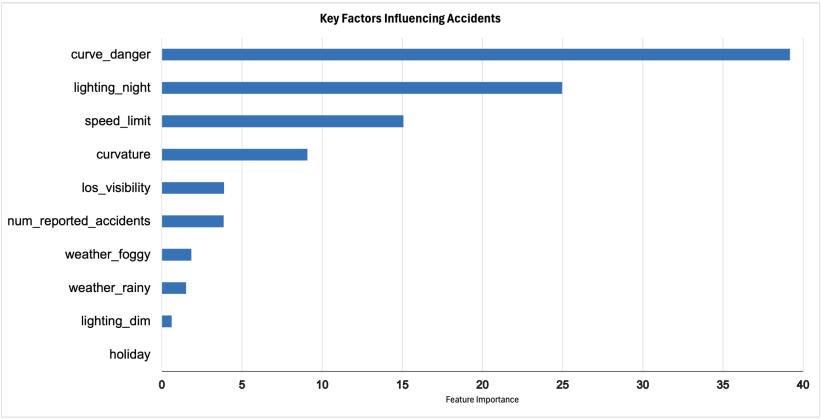



**Figure 3: Model Actual vs. Predicted Comparisons**

After applying the optimal parameters to our Random Forest model, its $R^2$ value increased by about 0.01 and its RSME decreased by about 0.025, which was a noticeable improvement compared to the untuned versions of each model. Figure 3 shows actual vs. predicted values for each model, where ideal models have points grouped as close to the 45 degree line as possible.

This model identified a number of features that were significant in predicting accident risk. The most important feature in the model was curve danger, which had an importance level of nearly twice any other feature. The other features that the model indicated were important were about the road conditions in regard to lighting and weather. The bar graph below shows the 10 features that the model selected and their respective importance level.




| Model Comparisons | | |
|---|---|---|
| **Model** | **R-squared** | **RMSE** |
| OLS | 0.8051 | 0.0735 |
| Random Forest | 0.8850 | 0.0564 |
| Gradient Boosting | 0.8813 | 0.0573 |
| Neural Networks | 0.8828 | 0.0569 |

**Figure 4: Feature Importance in Predicting Accident Risk & Model Comparisons**

## IV. Conclusion & Impact

Our predictive model directly answers the business questions by showing the state government where crashes are most likely, why they occur, and which interventions will have the greatest impact. It quantifies how curve severity, speed, visibility, weather, and holiday patterns influence accident risk, enabling proactive, data-driven decisions instead of reactive responses.

Curve danger emerged as the top risk factor, guiding speed limit reductions and road redesigns on sharp curves. Low visibility in rural areas directs targeted lighting improvements, while weather and holiday risks inform increased patrols and first responder readiness. These insights turn Safe System principles into actionable priorities, maximizing safety and reducing societal costs.

In practice, the model provides maintenance teams, transportation departments, and policymakers with a clear roadmap for resource allocation. RMSE confirms reliability, and with additional data such as granular weather, detailed road maps, and infrastructure records, accuracy and precision will improve. Our limitations include current gaps in environmental and geographic data, which require continuous updates and validation in a production environment.

## V. Proof of Kaggle Competition Submission