City of Austin Car Crashes

Luca Comba
University of St Thomas
St Paul, MN
comb6457@stthomas.edu

Hung Tran
University of St Thomas
St Paul, MN
@stthomas.edu

Steven Tran
University of St Thomas
St Paul, MN
@stthomas.edu

***Abstract*—After gathering the City of Austin's data set with over 216,000 reported car crashes from 2010, we were able to examine the relationship between various crash factors and their associated comprehensive costs using multiple machine learning approaches. Seven regression models, including Linear Regression, Ridge, LASSO, Decision Trees, Random Forest, Support Vector Regression, and K-Nearest Neighbors, were implemented to predict the estimated total comprehensive cost of accidents.**

The data set utilized is part of Austin's Vision Zero initiative and it includes features such as crash severity, vehicle types involved and the speed limit. Our analysis demonstrates that a simpler regression model can achieve the highest prediction accuracy with an R-squared value close to 1, providing valuable insights for urban planning and traffic safety improvements. The findings of this paper could contribute to understanding the economic impact of traffic accidents and can have an important impact on policy decisions for accident preventions.

***Index Terms*—austin, texas, car, crash, accidents, machine learning, linear regression, ridge, lasso, svr**

I. Introduction

The United States of America is heavily dependent on car transportation. Past data showed that "US cities in 1990 have levels of per capita auto use that are some two times higher than Australian cities [1]." Driving can be a dangerous and life-threatening action. A study conducted by the All India Institute of Medical Sciences (AIIMS) has shown that "the proportion of fatal accidents in total road accidents has consistently increased since 2002 from 18.1% to 24.4% in 2011 [2]."

With the advent of new technologies and the publication of new open source machine learning libraries such as Scikit-learn which is "a machine learning package in the Python programming language that is widely used in data science [3]", it has been easier to analyze publicly available data set.

The City of Austin, Texas (U.S.A.), in the last few years, has started a new city project to better understand and prevent car accidents. The Austin Transportation program has worked on the "Vision Zero" project to create a transportation network that protects human life. The intent of the project is to design a city network that focuses on safety for all of its citizens [5]. Our research project will heavily rely on the Vision Zero project and on its powerful data set that will be discussed in the next section. The research effort aims at empowering the data collected by the City of Austin, and better understand the current situation of car accidents as well as focusing on predicting and understanding the Estimated Total Comprehensive Cost of a car accident.

A. The Austin Crash Report

The Austin Crash Report Data Set [6] includes records of traffic accidents in the city of Austin, Texas, from 2010 to today, with 216,088 instances and 45 features, including both numerical and categorical data. The data set is managed by the Texas Department of Transportation, and utilized by Crash Record Information System database, which is populated by reports submitted by Texas Peace Officers throughout the state, including Austin Police Department.

The original data is composed by 45 features. In general terms, the data set had information about aspects such as location (latitude / longitude), crash timestamp, speed limit, severity (0-5), injury and death counts, type of vehicle involved, and the estimated total comprehensive cost. Many of these features will be used for training the machine learning models.

B. Estimated Total Comprehensive Cost

The Austin's Vision Zero initiative quantifies the impact of traffic crashes using comprehensive costs, which integrate both economic factors such as medical expenses, lost wages. The comprehensive cost also monetized quality of life impacts based on national guidelines. This total societal cost per crash is derived from National Safety Council and Federal Highway Administration methodologies [7].

These comprehensive crash costs are a critical tool for Austin's transportation planning, enabling a data-driven approach to understanding and mitigating traffic-related harm. In our research, we focused on predicting the Estimated Total Comprehensive Cost as a key metric. This predicted value allows us to analyze its relationship with various factors, such as posted speed limits and vehicle types involved, to inform future safety interventions and policy decisions.

C. Data Exploration

An exploratory data analysis was conducted on the The Austin Crash Report Data Set previously introduced. After an initial examination of the data types and a confirmation of no missing values a basic descriptive statistics such as mean, median, and standard deviation were computed for key cost-related variables, specifically on the estimated_total_comprehensive_cost, to understand their central tendencies and spread. It was discovered that the total_comprehensive_cost mean was of 307,980$.

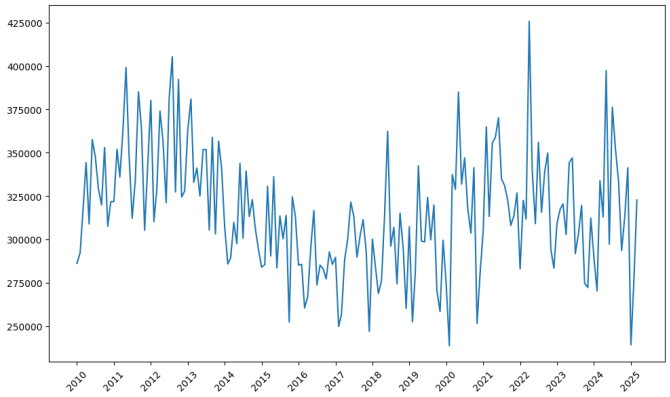

Fig. 1. Total Comprehensive Cost by month

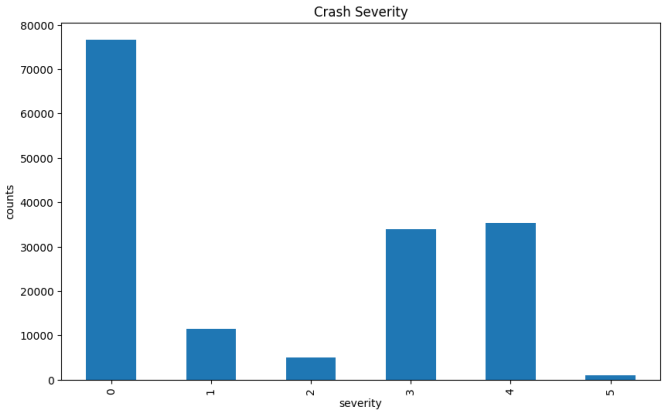

Fig. 2. Crash Severity counts

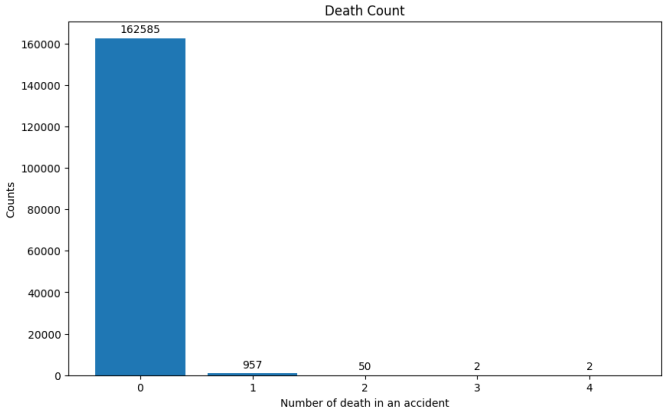

Fig. 3. Crash severity counts. From the least severe (0) to the most severe crashes (5)

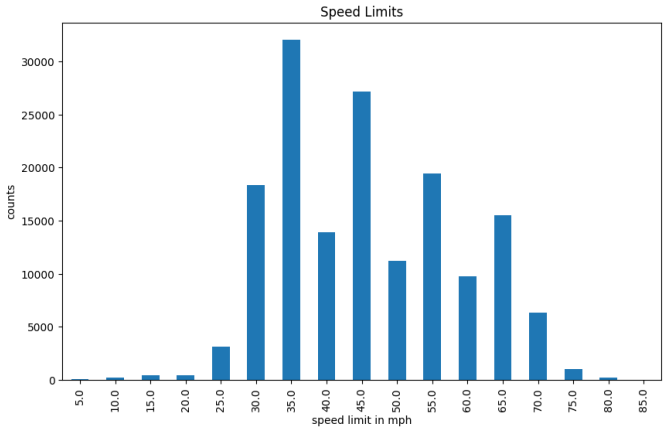

Fig. 4. Speed limits represented in the data set in mph

It was then discovered that most of the crashes resulted in a non-significant severity, and in a non fatal crash. It is calculated that the data set has a 0.65% death rate of all accidents (Fig. 3).

The distribution of the speed limits contained in the data set shows a normal distribution, centered around 45 miles per hour (Fig. 4).

II. Data Cleansing

The data set utilized underwent a comprehensive, multistage cleaning process to ensure data quality and suitability for subsequent analysis.

The initial phase of the data cleaning focused on several data transformations. First, records flagged as temporary or deleted were removed from the original data set. Following this, a significant number of columns were determined to be redundant or not conducive to the modeling process and were consequently dropped. This included various identifier columns such as 'ID', 'Crash ID', and 'case_id'. Columns containing repetitive address details such as 'rpt_block_num', 'rpt_street_name', 'rpt_street_sfx'. A 'point' column derived from geographic coordinates, and the 'Reported street prefix' were also removed. Additionally, columns pertaining to road type like the 'private_dr_fl', 'onsys_fl' columns were excluded as initial exploration revealed no instances of accidents on private roads. After deriving a new 'Crash timestamp (US/Central)' from the original column 'Crash timestamp', the latter was dropped. The 'estimated_maximum_comprehensive_cost' was removed from the feature set as it would be a repetition of the predictive variable which we have set to the Estimated Total Comprehensive Cost.

To have consistent data types and appropriate formatting we applied some other data transformation. Column naming conventions were standardized by converting all names to lowercase and replacing spaces with underscores ('Primary address' became 'primary_address'). The 'Crash timestamp (US/Central)' column was converted into a datetime object to facilitate time-based feature engineering. Further, several columns received more descriptive names; for example, 'crash_fatal_fl' was changed to 'fatal_crash', 'crash_speed_limit' to 'speed_limit', and 'crash_sev_id' to 'crash_severity'.

Additional value corrections and transformations were needed for the easy of use by the future machine learning models. For the 'speed_limit' feature, values of -1 and 0, likely representing missing or erroneous data, were replaced with

NaN and subsequently removed. The remaining 'speed_limit' values were then binned by rounding them down to the nearest multiple of 5. This was done for helping with a data visualization effort.

The 'crash_severity' column was remapped to an ordinal scale (0=NOT INJURED, 1=UNKNOWN, 2=POSSIBLE INJURY, 3=NON-INCAPACITATING INJURY, 4=INCAPACITATING INJURY, 5=KILLED) to represent it from the lowest to the highest value. These remapped categories were then one-hot encoded to create individual binary features for each severity level.

The 'units_involved' column, a textual description of vehicles in an accident, was also extensively processed. Combinations of units appearing infrequently, occurences of less than 1000 were filtered out. Unique vehicle types were then parsed from these strings, and new binary indicator columns were generated for each type, such as 'unit_involved_bicycle'. The original 'units_involved' column was then discarded.

Furthermore, the 'timestamp_us_central' feature was leveraged to engineer new time-related attributes, including 'hour', 'day_of_week', 'month', 'year', 'day_of_month', and a binary 'weekend' indicator. To capture the cyclical nature of time, sine and cosine transformations were applied to the 'hour' and 'month' features, creating 'hour_sin', 'hour_cos', 'month_sin', and 'month_cos'.

Finally, any remaining rows containing NaN values after these steps were removed to ensure a complete data set.

Our final machine learning models did not utilized other columns that were kept after the cleaning of the data set. The unutilized columns were: 'primary_address', 'secondary_address', 'timestamp_us_central', 'latitude', and 'longitude'. This decision was made due to their data types or their deemed relevance for the specific modeling approach chosen. Although geo-spatial features like latitude and longitude could have been used for creating location clusters, they were excluded in this analytical iteration for the need of simpler models.

Lastly, for modeling purposes, the target variable, 'estimated_total_comprehensive_cost', was scaled by a factor of 100,000. The previously described cleaning and preparation stages resulted in a refined dataset optimized for the subsequent feature selection and model development processes.

III. Feature Selection

Given our dataset has 45 features, some might be redundant and irrelevant, potentially causing a longer computation time and reducing the efficiency of the models. To make sure our model concentrates on only useful independent variables, we select backward elimination to remove nonsignificant features that have a p-value under 0.05, avoid overfitting, and increase the overall performance of all models.

Backward Elimination

Overview

This method will start with the full 47 features in the original data set and remove the higher p-value feature in each iteration until all remaining features are statistically significant. By following this approach, our dataset will left with only relevant features to use for predictive purpose.

Result

Here is a list of 19 features that have been eliminated:

- month
- construction_zone
- unit_involved_large_passenger_vehicle
- month_sin
- day_of_month
- unit_involved_pedestrian
- day_of_week
- severity_possible_injury
- severity_incapacitating_injury
- severity_non_incapacitating_injury
- law_enforcement_fatality_count
- unit_involved_bicycle
- year
- death_cnt
- tot_injry_cnt
- unit_involved_passenger_car
- month_cos
- hour_sin
- unit_involved_other_unknown

IV. Feature Scaling

A. Target Variables

As our input variables are mostly in the range 1 to 10, we decided to divide our target variables (Estimated Total Comprehensive Cost) by 100,000 to ensure numerical stability and potentially help improve computation efficiency.

B. Independent Variables

We will apply Standardization to center the numerical data around the mean of 0 and the standard deviation of 1. Boolean columns will remain unchanged.

V. Methodology

To predict the estimated total comprehensive cost, we will use 7 different machine learning models (listed below). In order to improve consistency and code organization, we will also leverage a pipeline function in sklearn. It allows preprocessor settings, which will be helpful in our later step of hyperparameter tuning.

A. Linear Regression

This method will capture a linear relationship, and by comparing the coefficients, we can evaluate how each feature impacts the target variables and by how much.

B. Ridge Regression

This method will work well with a multiple-feature dataset and can penalize non-significant ones, closing the gap between training and testing errors, leading to overfitting prevention.

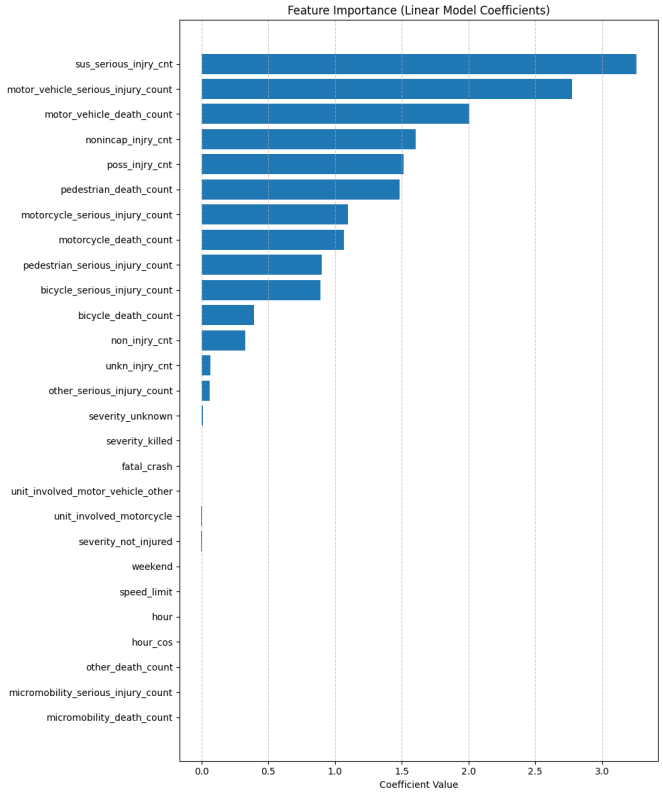

Fig. 5. Weight of each Linear Regression model's coefficient.

C. LASSO Regression

This method will apply L1 regularization, automatically do feature selection by shrinking the coefficients of non-significant features to zero, helping to reduce overfitting.

D. Decision Tree Regression

This method will capture non-linear relationships and won't be impacted by outliers while predicting the cost.

E. Random Forest Regression

This method combines multiple decision trees to enhance the models' accuracy.

F. Support Vector Regressor (SVR) Regression

This method will use kernels to handle non-linear data, transform our dataset to high-dimensional spaces, and then choose the best hyperplane to reduce prediction errors.

G. K-Nearest Neighbors Regressor (KNN) Regression

This method will apply a non-parametric approach to capture complex relationships, and then use only local variations to predict the cost.

VI. Results

A. Linear Regression

Figure 5 presents the coefficient values of features used in the models. Temporal variables such as accident timing (e.g., weekend, hour) and vehicle type (e.g., motorcycle, car) exhibited relatively low weights. In contrast, the number of serious injuries and fatalities emerged as the most influential predictors, indicating their strong association with the model's output.

B. Models accuracy scores

Figure 6 presents the R-squared values for both the training and testing sets across all regression models. These values reflect how well each model explains the variance in the target variable. All models, with the exception of Support Vector Regression (SVR), achieved exceptionally high R-squared scores ranging from 0.9 to 1. In contrast, the SVR model yielded significantly lower R-squared values of 0.88 on the training set and 0.83 on the test set, indicating poor explanatory performance. The near-perfect scores observed in the other models raise concerns about potential overfitting.

Figure 7 presents the Root Mean Square Error (RMSE) for each model. The linear regression model, along with its ridge variant, had the lowest RMSEs across both sets. These extremely low error rates may indicate data leakage, suggesting overlap between training and testing data. In contrast, the decision tree, random forest, and k-nearest neighbors regressors showed large increases in RMSE from training to testing sets.

These spikes indicate overfitting and weak generalization to unseen data.

Overall, linear regression performed best in terms of RMSE. The minimal gains from Ridge and Lasso suggest limited multicollinearity in the features, reinforced by the fact that GridSearchCV identified very low optimal alpha values (0.1) for both. However, the unusually high R^2 and low RMSE values highlight a need to investigate potential data leakage.

For future work, the following actions are recommended: carefully verify data splitting to rule out leakage; combatting overfitting by expansion of hyperparameter search spaces for tree-based and instance-based models, such as increasing the number of trees, reducing tree depth, and raising the minimum number of samples per split; and consider applying Principal Component Analysis (PCA) to reduce feature complexity and improve generalization.

C. Predictions

After the training and test of the machine learning models previously described, we have selected a random observation from the data set to see the resulting predictions for each model.

The randomly selected observation was characterized by having the features of not being a fatal crash and with a speed limit of 55. The observation had a true observed Estimated Total Comprehensive Cost of 140,000.

When each model was asked to predict the Estimated Total Comprehensive Cost, a multitude of results were generated. As previously mentioned, the Linear Regression and the Ridge Regression had a good accurate prediction, as they resulted in a Estimated Total Comprehensive Cost of 139,000. Also

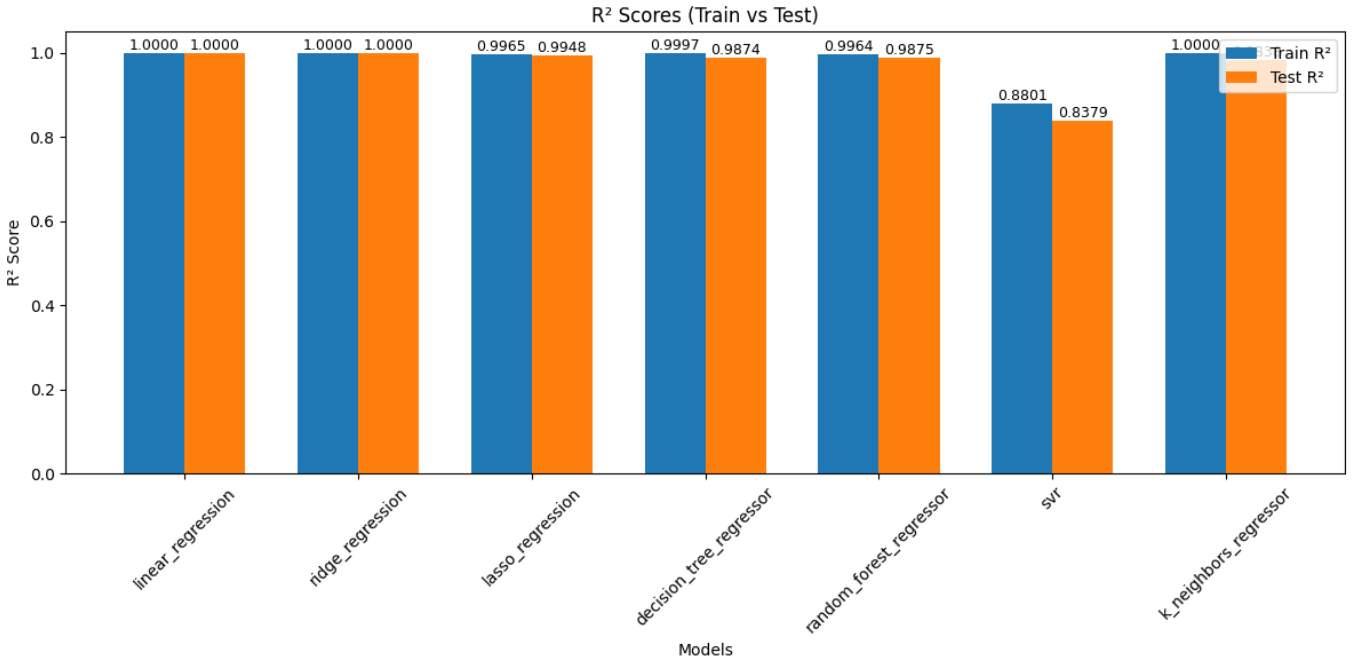

Fig. 6. R^2 score of each model trained.

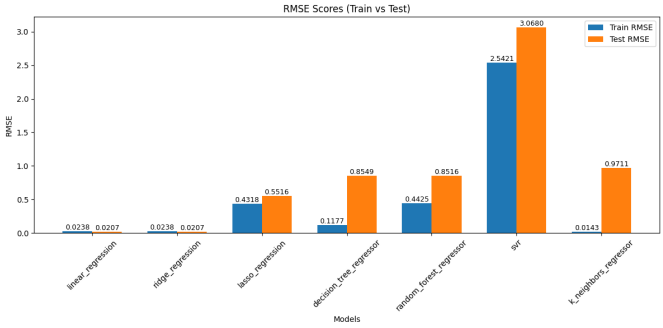

Fig. 7. RMSE score of each model trained.

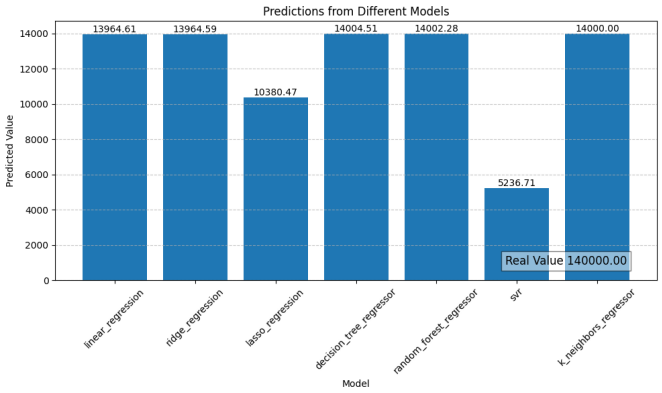

Fig. 8. Predictions of the Comprehensive Cost using a observation from the data set. The true value of the Comprehensive Cost in the data set was of 140,000$.

the Decision Tree, the Random Forest Regression and the K Neighbors Regression correctly predicted a value of 140,000. The Lasso Regressor and the SVR models incorrectly predicted the already seen observation (Figure 8).

These predictions confirm the scores seen previously, as the SVR had the lowest R^2 value of all models and resulted in an incorrect prediction.

After testing our models with a observation, which they might already seen, we decided to generate a new observation and set the observation to be a fatal crash, with a speed limit of 80 miles per hour and a severity of killed.

The true Estimated Total Comprehensive Cost for the new observation will remain unknown, although our trained models calculated their prediction. The resulting predictions were sparse. The Linear Regression and the Lasso Regressor both predicted the Estimated Total Comprehensive Cost to be 14,000, while the Lasso Regressor model, the Random Forest Regressor and the SVR predicted the value to be over 120,000 (SVR was the highest value of 219,000). The Decision Tree Regressor and the K Neighbors Regressor models had the lowest Estimated Total Comprehensive Cost prediction of less than 5,000.

These prediction tell us that the models do not do a good job at predicting at a new observation, because we would have been expecting to have a higher Estimated Total Comprehensive Cost for a higher car accident severity.

VII. Conclusion

This research project worked on predicting the Estimated Total Comprehensive Cost of car crashes in Austin, Texas, by

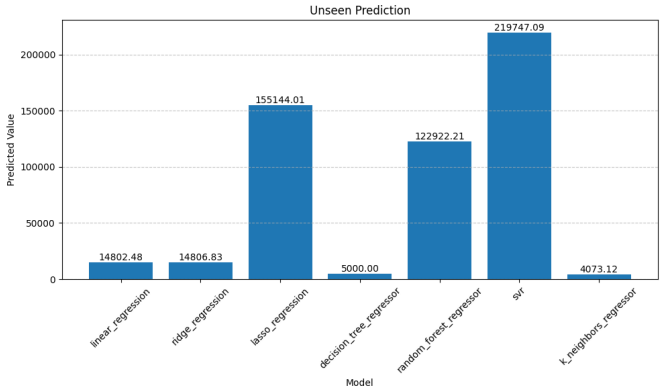

Fig. 9. Predictions of the Comprehensive Cost unseen already variables.

applying seven distinct machine learning regression models to a comprehensive dataset from the City of Austin's Vision Zero initiative. Our analysis revealed that simpler models, particularly Linear Regression and its regularized variant Ridge Regression, can hold the highest predictive accuracy.

Feature importance analysis from the Linear Regression model indicated that the number of serious injuries and fatalities were the most significant predictors of crash cost, while temporal factors and vehicle types had lesser impact.

Although we have successfully predicted some car crash observation, our machine learning models can be improved. A better review of the data and features might be needed as the high R^2 score of the machine learning model could demonstrate the models to be over fitted. As previously mentioned PCA and feature reduction might solve some of the overfitting problems.

Despite these limitations, this research provides a foundational step towards leveraging machine learning for enhanced traffic safety in Austin. Improving the robustness of these machine learning models, and thanks to future iterations the research space can offer more reliable insights for policy decisions and traffic accident prevention.

REFERENCES

[1] J. R. Kenworthy and F. B. Laube, "Automobile dependence in cities: An international comparison of urban transport and land use patterns with implications for sustainability," Environmental Impact Assessment Review, vol. 16, no. 4–6, pp. 279–308, Jul. 1996, doi: https://doi.org/10.1016/s0195-9255(96)00023-6.

[2] M. Ruikar, "National statistics of road traffic accidents in India," Journal of Orthopedics, Traumatology and Rehabilitation, vol. 6, no. 1, p. 1, 2013, doi: https://doi.org/10.4103/0975-7341.118718.

[3] J. Hao, "Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language," Journal of Educational and Behavioral Statistics, vol. 44, no. 3, p. 107699861983224, Feb. 2019, doi: https://doi.org/10.3102/1076998619832248.

[4] "Data.gov," Data.gov, 2024. https://catalog.data.gov/dataset/vision-zero-crash-report-data

[5] "Vision Zero Viewer" visionzero.austin.gov. https://visionzero.austin.gov/viewer/

[6] City of Austin Texas, "Austin Crash Report Data - Crash Level Records," Austintexas.gov, Jul. 30, 2019. https://data.austintexas.gov/Transportation-and-Mobility/Austin-Crash-Report-Data-Crash-Level-Records/y2wy-tgr5/about_data

[7] "Comprehensive Crash Costs — AustinTexas.gov," Austintexas.gov, 2018. https://www.austintexas.gov/crashcosts