California Traffic Collision Data - Analysis Final Report

Introduction

A data set containing every recorded traffic collision between 2001 and 2020 in Los Angeles was used to investigate contributing factors that lead to increased collision frequency and severity. This data was originally collected from the California Highway Patrol. Our focus is on Los Angeles County, which had approximately 2.85 million recorded accidents. Our dataset originally contained 75 variables but this will be subsetted depending on the analysis completed, which are outlined below.

Our exploratory data analysis identified alcohol involvement and time of day as key contributors to traffic collisions. The analysis below aims to explore these factors further and also determine additional significant predictors in traffic collisions. It was determined that the following investigations would be conducted:

- Fitting Hour of day versus P(alcohol use): We will attempt to fit a distribution to model the proportion of collisions involving alcohol by time of day
- Distributions of the number of injuries per collision, alcohol vs. no alcohol: We will determine distributions to fit the number of injuries per collision with and without alcohol involvement. We will then compare these distributions to determine if alcohol does have an impact on the number of injuries
- Time Series Forecasting Analysis: We will create a model trained on data up to 2019 and use this to predict the number of collisions in 2020. We will compare the actual data in 2020 to determine the impact of the COVID-19 pandemic on the number of collisions. We will also conduct the same comparison for alcohol related collisions.
- Random Forest Regression: We will use a Random Forest model to determine the significant predictors in collision severity
- Lasso Regression: We will use a Lasso Regression model to determine the significant predictors in collision severity

By conducting these analyses, we aim to gain a deeper understanding of how alcohol use and other contributing factors affect the incidence rate and severity of traffic collisions in Los Angeles County.

Methods

Fitting Hour of day versus P(alcohol use)

Initially, distributions of the number of collisions by hour of day were compared across whether alcohol was involved or not (see Figure 3). These showed clear differences and it was decided to explore the probability of alcohol involvement for each hour of the day.

The `collision_time` column was coerced to an `hour` column, and then data were aggregated by `hour` and the proportion of collisions involving alcohol calculated for each hour. A smooth trend was observed and a B-spline was fitted to the data (see Figure 4).

Distributions of the number of injuries per collision, alcohol vs. no alcohol

We wanted to investigate if alcohol had an impact on the number of injuries during collisions. This would be one metric for measuring if the severity of collisions increased when alcohol was involved. The plot below shows the distribution of the number of injuries with alcohol and without. Initially they look quite similar and it appears that alcohol may have no impact on the number of injuries.

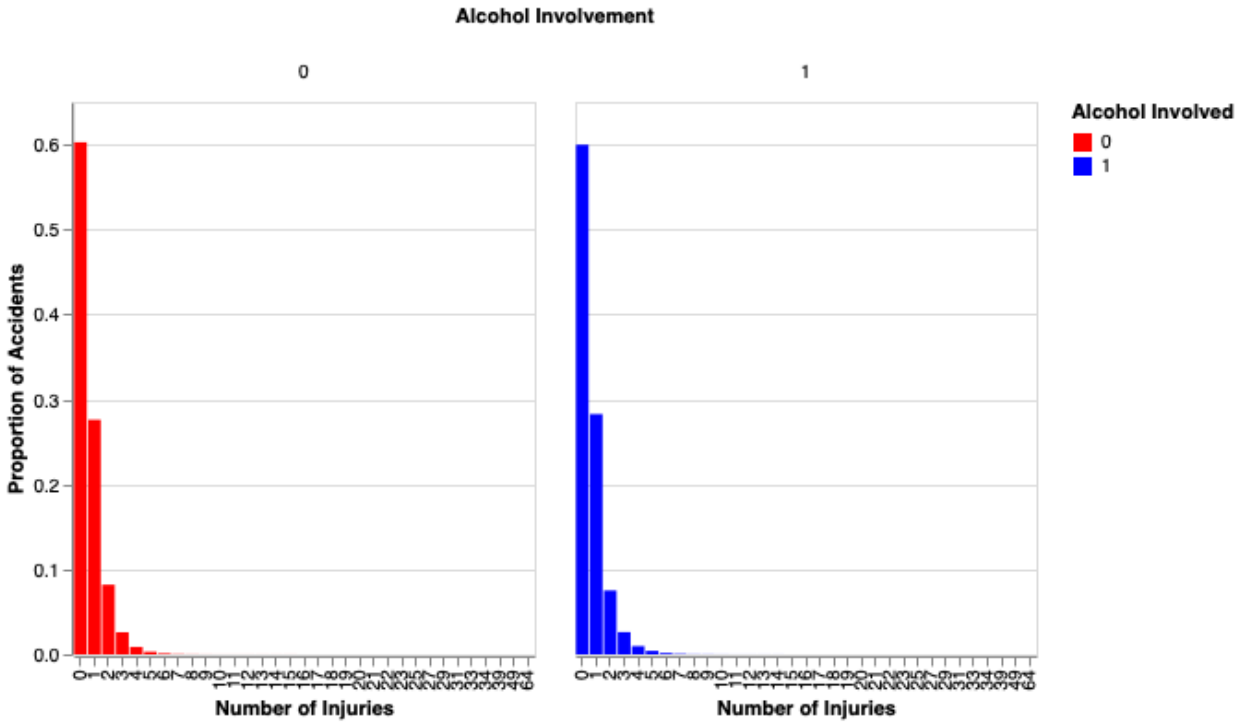

Figure 1: Injury frequency relative to the total number of accidents

To investigate this further, we decided to fit models to each of these subsets to determine if there was a statistical difference between alcohol and no alcohol involvement in collisions. Exponential and Weibull distributions were fit to these subsets of data and the log likelihood of each distribution was calculated. In both cases the Weibull distribution was the better fit.

We wanted to determine if there was an impact on subsetting based on alcohol so we repeated the process for the dataset without discriminating based on alcohol use. An exponential and Weibull distribution was fit to the entire dataset and based on log likelihood values we determined that the Weibull distribution was again the better fit. In order for this to be an accurate comparison, the log likelihood of the alcohol and non alcohol groups were added together to get the log likelihood of the entire dataset but based on two separate models. We performed a likelihood ratio test on the two distributions and found that there was in fact a difference. This means that alcohol does have some impact on the number of injuries per collision.

Time Series Forecasting Analysis

Upon analyzing the annual collision data, we observe a significant decline in accidents in 2020 compared to previous years. This is unsurprising given the impact of the COVID-19 pandemic. However, the key question is: how large is the discrepancy between the observed and the expected number of accidents?

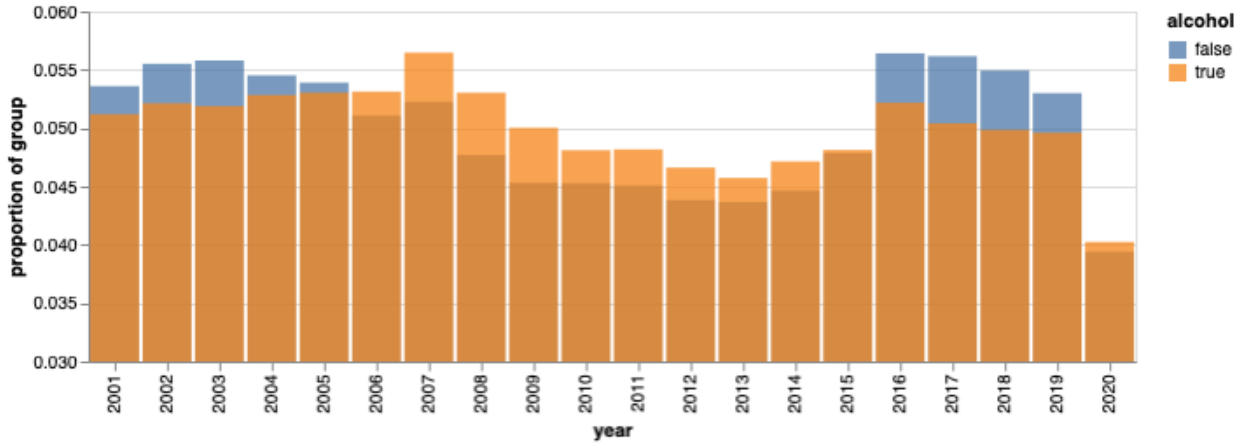

Figure 2: Total accidents over time by alcohol involvement. The y-axis is scaled to the total number of accidents in each group

To address this, we utilize a time series model to forecast the number of accidents in 2020 based on historical data from previous years. Specifically, we have chosen the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model, as it is well-suited to account for seasonal patterns and trends, unlike the standard ARIMA model. Before modeling, we performed data preprocessing, aggregating the data on a monthly basis to better capture the underlying trends. Monthly aggregation facilitates finer-grained model training and provides clearer insights into seasonal patterns, such as higher collision rates during summer months. This should, in theory, enhance the model's ability to predict 2020 data, while also enabling the detection of month-to-month fluctuations during the pandemic period. The data was also subsetted to only include the collision data and because each row represents a collision, no further variables were needed for the analysis.

For the ARIMA parameters, we have opted for the default values of 1 for each, as the data does not exhibit complex seasonal dependencies. The seasonal parameters have been set with p=1, meaning we are using the data from the same month in the previous year for prediction, and s=12 to account for annual seasonality. We have chosen 12 steps for the forecast, corresponding to the 12 months of 2020, which is the time period we are forecasting for.

Random Forest Regression

To prepare the data set for Random Forest (RF) regression, the initial data set with all predictors was considered. The collision time feature was converted to three numerical columns: year, day (of year), and minute (of day). A series of predictors were dropped for logical reasons. Then, predictors where > 80% of values were missing were dropped. Finally, remaining categorical predictors were removed if they had more than 100 different categories. Of the 75 variables in the initial set, 43 remained after the pre-processing steps.

jurisdiction	chp_shift	population	county_city_location	county_location
special_condition	beat_type	chp_beat_type	chp_beat_class	distance
direction	intersection	weather_1	state_highway_indicator	location_type
side_of_highway	tow_away	party_count	primary_collision_factor	pcf_violation_category
pcf_violation_subsection	hit_and_run	type_of_collision	motor_vehicle_involved_with	pedestrian_action
road_surface	road_condition_1	lighting	control_device	chp_road_type
pedestrian_collision	bicycle_collision	motorcycle_collision	truck_collision	not_private_property
alcohol_involved	statewide_vehicle_type_at_fault	chp_vehicle_type_at_fault	latitude	longitude
year	minute	day		

Table 1: List of predictors used for Random Forest model fitting

The pre-processed data were one-hot encoded and partitioned into test and training sets (20% test size). A Random Forest model was fitted to the training data against the `injured_victims` label using the scikit-learn `RandomForestRegressor` class. Default parameters were used except for `n_estimators` which was reduced to 10 due to the computational intensity of this task. Using this model against partitioned test data returned predictions with root mean squared error of 0.785 (in units of injured victims) and returned the correct number of injured victims (after rounding predictions to integer values) for 68.4% of the test data.

The above procedure was repeated but using the ratio of `injured_victims` over the total `party_count` as a “proportion of injuries to involved parties” label. (Note that `party_count` was removed from the features used to test and train here.) This fit returned RMSE of 0.406 (in units of injured victims per involved party) with an accuracy of 60.6%.

Finally, the `RandomForestClassifier` class was used to similarly fit the training data, this time to predict the categorical `collision_severity` label. This model correctly predicted the category for 68.1% of the test data.

For the three fitted Random Forest models, feature importances were compared (see Table 3).The importance of `minute` and `hit_and_run_misdemeanor` were further explored through visualizations (see Figures 9 and 10).

Lasso Regression

We chose Lasso for this regression task because it performs feature selection and shrinks some of the coefficients of predictors to 0. This effectively removes irrelevant features. Pre-processing the data was completed using the same steps described in the Random Forest regression methodology. The only difference is that Lasso does not accept null values in the data so these were filled using the mean for each column. Also, we had to choose a subset of the data due to computation power. We could not model across all years of the data due to its size. We chose to subset the data to the year 2020 since this was the most recent data. This was done under the assumption that 2020 would be a representative sample for the entire dataset. This would need to be verified at a later time but due to computational and time constraints we were not able to do so.

Similar to Random Forest, we first fitted to the `injured_victims` label. This returned predictions with a root mean squared error of .853. It determined that the following predictors were significant: `hit_and_run_misdemeanor, party_count, minute, distance`. We again repeated this using the ratio of `injured_victims` over the total `party_count` as a "proportion of injuries to involved parties" label. This returned predictions with a root mean squared error of 0.438 and determined that only `distance` was relevant. The `collision_severity` label is categorical so we did not perform Lasso regression on this label.

Discussion

Fitting Hour of day versus P(alcohol use)

A stark difference was seen in the collision by hour distributions when separated into those involving alcohol and those not:

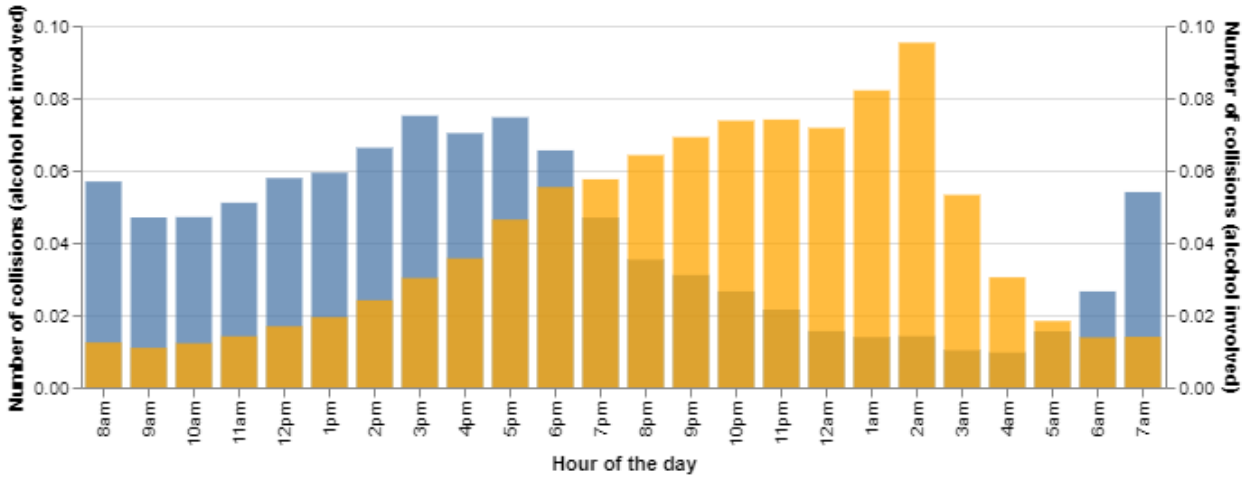

Figure 3: Distributions across hours of the day of collisions involving alcohol (orange) versus those not involving alcohol (blue). Note that the y-axes have been scaled by the number of collisions in each group to make distributions more easily comparable. Overall 8.5% of total collisions involved alcohol.

Differences in these distributions are obvious. Collisions without alcohol involved have two local maxima presumably coinciding with morning and evening commutes, whereas alcohol-involving collisions peak late at night.

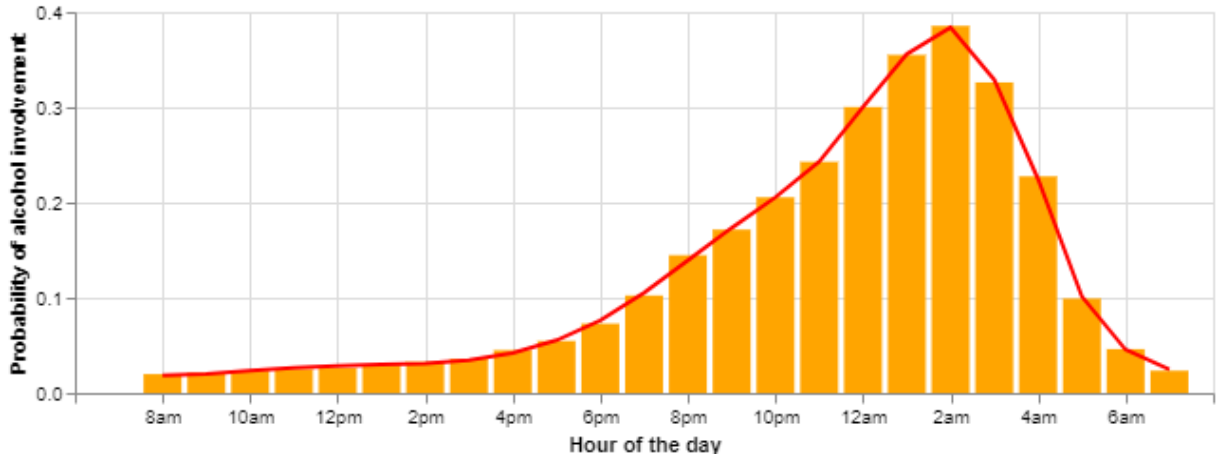

Figure 4: Probability of alcohol involvement in a collision per hour of the day. A B-spline (red) is shown fitted to these data.

An interesting observation in these plots is the spike seen at 2am in the number of collisions involving alcohol – this stands out in an otherwise smooth, nearly-normal distribution. This coincides with the time of maximum probability of alcohol involvement. The noticeable jump in collisions might be related to the increase in traffic from bars after last call – 2am is when LA bars legally must stop serving alcohol.

Distributions of the number of injuries per collision, alcohol vs. no alcohol

In the methodology section above we determine that there was in fact a statistical difference between the distribution of the number of injuries per collision with and without alcohol involvement. We will now look at the parameters of these weibull distributions to determine the relationship alcohol has with the number of injuries.

Alcohol-Involved Collisions:
Shape parameter (α): 0.2828
Scale parameter (β): 1.0 (approximately)
Location parameter (γ): 32,408.77

Non-Alcohol-Involved Collisions:
Shape parameter (α): 0.1101
Scale parameter (β): 1.0 (approximately)
Location parameter (γ): 175,746.96

A smaller shape parameter indicates that the distribution has a "longer tail," meaning that accidents with a higher number of injuries are less likely but still possible. For

alcohol-involved accidents, the shape parameter is higher (0.2828) compared to non-alcohol-involved accidents (0.1101), suggesting that alcohol-involved accidents have fewer collisions with a large number of injuries. The scale parameter was nearly identical for both distributions. The location parameter shifts the distribution along the x-axis. For non-alcohol accidents, the location parameter is much higher (175,746.96) than for alcohol-involved accidents (32,408.77). This suggests that collisions without alcohol tend to have a higher number of injuries per collision.

In summary, there was a difference between the distributions of the number of injuries per collision when subsetting for alcohol involvement. However, it did not follow the assumptions we had going into this analysis. Due to the well known danger associated with drinking and driving, it was our assumption that the number of injuries would have increased when alcohol was involved. It is in fact the opposite. The non-alcohol collisions had a longer tail and were more likely to have a higher number of injuries per collision. This unexpected result could be related to our previous finding – perhaps alcohol-related crashes have fewer expected injured victims because they more commonly happen in the middle of the night rather than at peak traffic hours.

Time Series Forecasting Analysis

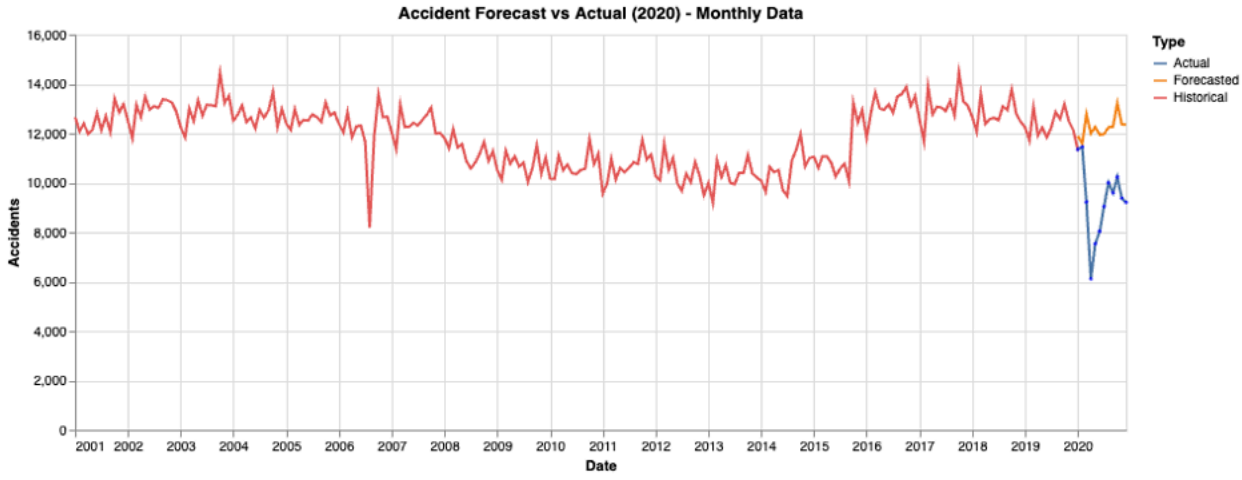

Figure 5: Historical number of monthly accidents, and forecasted number of monthly accidents compared to actual in 2020

After training the model and comparing the forecasted values with the actual 2020 data, we observed a notable discrepancy, which was anticipated. Upon closer inspection of Figure 5, it is evident that the predicted values for January and February closely align with the actual data. However, in March and April, there is a significant gap between the forecasted and actual values. This period corresponds to the strict lockdowns that

resulted in fewer vehicles on the road, leading to fewer collisions. This shift in behavior is an anomaly that the SARIMA model, which relies on historical patterns, cannot easily account for. Notably, the forecasted values dip slightly during the summer months, following typical trends observed in previous years. Would we observe a similar discrepancy if we separate the data into alcohol-related and non-alcohol-related accidents?

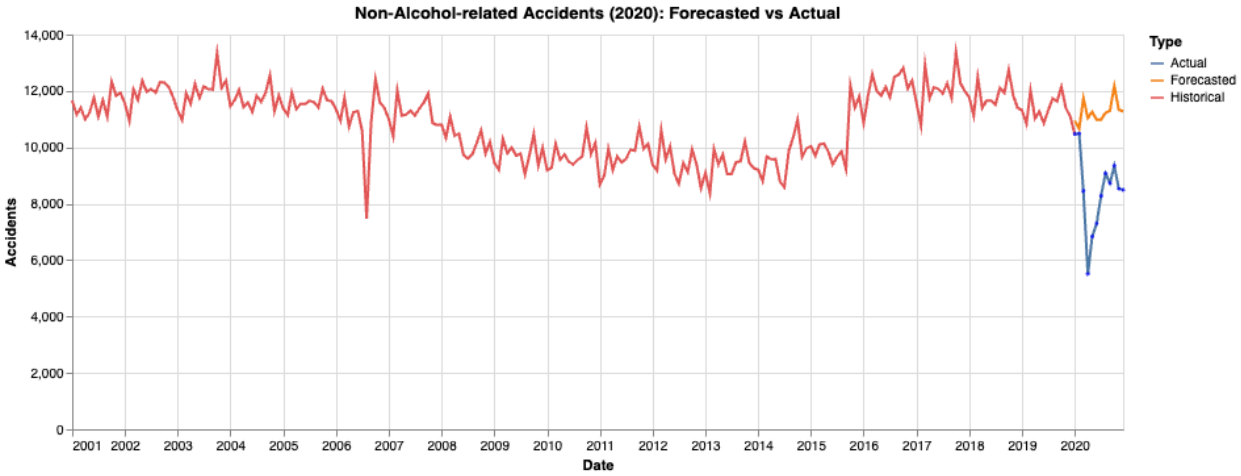

Figure 6: Historical number of monthly non-alcohol related accidents, and forecasted number of monthly non-alcohol related accidents compared to actual in 2020

Unsurprisingly, the trend for non-alcohol related accidents closely resembles Figure 5, as non-alcohol related accidents make up about 90% of the data. However, we do observe a similar pattern in alcohol-related accidents, but the magnitude of the difference is much smaller. In fact, the forecast for alcohol-related accidents comes very close to the actual values, with a mere 74-accident difference in August (1,000 predicted vs. 926 actual). This suggests that while there is still a decline in alcohol-related accidents compared to the forecast, the difference is less pronounced than in non-alcohol-related accidents. This implies that alcohol-related accidents exhibit more predictable trends and are less influenced by external factors, such as changes in traffic volume, which heavily affected non-alcohol-related accidents during the pandemic.

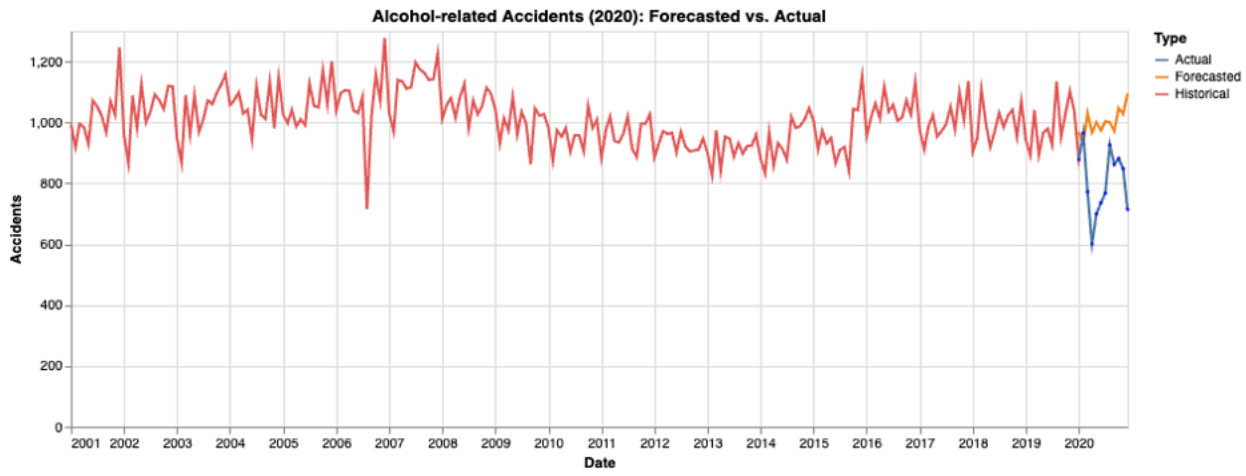

Figure 7: Historical number of monthly alcohol related accidents, and forecasted number of monthly alcohol related accidents compared to actual in 2020

The SARIMA model, by its nature, relies on historical data and assumes that future patterns will closely resemble those of the past. Therefore, the patterns of alcohol consumption and drinking-and-driving behavior may have remained relatively stable throughout the pandemic, particularly as the lockdown restrictions eased later in 2020.

It is important to note that a model like SARIMA may struggle with forecasting during an outlier year like 2020, given the unprecedented disruptions caused by the pandemic. For predictive accuracy in such cases, more sophisticated models may be necessary. However, the goal of this analysis was not solely to achieve precise predictions, but to evaluate trends and explore the differences between alcohol-related and non-alcohol-related accidents in the context of the pandemic.

Random Forest Regression

The initial data pre-processing involved modifying some features. Although `hour` (of the day) was a useful feature for previous analysis, in this case it was not used. Instead, `year`, `day` (of the year), and `minute` (of the day) were used. These capture the timestamp information down to minute resolution, and serve to be more readable than a timedelta value, but less redundant than more granular time feature columns would be (*e.g.* all of year, month, day, hour, minute).

Several predictors were also removed altogether prior to RF regression. Some were removed as they were expected to obviously correlate with our labels (*e.g.* using `severe_injury_count` to predict `injured_victims`). Others were removed based on reasonable expectations that they wouldn't help the model, such as the unique

`case_id` assigned to every collision. All features where more than 80% of rows were missing values were removed.

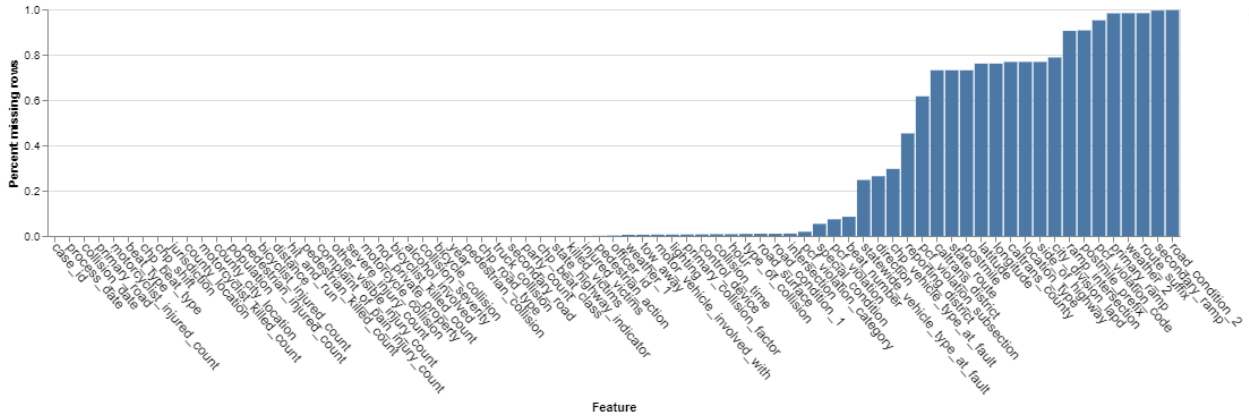

Figure 8: The proportion of rows missing for each feature remaining in the data set at this step. The 8 rightmost features in this figure had more than 80% missing fields and were removed from the analysis.

The removal of columns with many missing rows was *not* done to improve model performance or interpretation. In this data set, some important fields (such as `alcohol_involved` are coded such that the affirmative is 1, and the negative is `NaN`. Recklessly removing missing data compromises the analysis by introducing undue bias; for several columns in this data set, missing values are meaningful. This is why a relatively high threshold of 0.8 was used, and the individual columns were examined before removal. These columns were removed for practical purposes: it was suspected that columns with such a high missing data rate were not in regular use, and so would not be practically useful features to deem important later on. For example, some of these (such as `pcf_violation_code`) are defunct fields used only by a specific department, and some (such as `secondary_ramp` or `road_condition_2`) are presumably secondary “overflow” fields that appear to be rarely used by police officers.

column	unique
secondary_road	109860
primary_road	94213
officer_id	51563
beat_number	33877
reporting_district	11534
county_city_location	89

Table 2: The top 6 features by unique value count. The first 5 were removed, `county_city_location` was retained.

The fitted Random Forest models' feature importances were examined:

feature	importance
minute	0.097903
jurisdiction_1942	0.086192
hit_and_run_misdemeanor	0.079744
distance	0.075508
year	0.074853
tow_away_1	0.049616
longitude	0.029782
motor_vehicle_involved_with_other motor vehicle	0.028568
latitude	0.028077
party_count	0.024574

feature	importance
minute	0.103374
distance	0.083997
hit_and_run_misdemeanor	0.078663
year	0.077538
jurisdiction_1942	0.071948
tow_away_1	0.053825
longitude	0.031742
latitude	0.030364
motor_vehicle_involved_with_parked motor vehicle	0.013555
type_of_collision_sideswipe	0.012763

feature	importance
minute	0.105863
year	0.090825
distance	0.069621
hit_and_run_misdemeanor	0.047007
tow_away_1	0.022380
latitude	0.022078
longitude	0.022059
hit_and_run_felony	0.021800
party_count	0.020618
tow_away_0	0.018855

Table 3: Top ten predictors from feature importance of Random Forest models fitted against `injured_victims` (left), `injured_victims` / `party_size` (middle), and `collision_severity` (right)

The comparison of feature importances led to an investigation of injuries per hour of the day. Although `minute` was the variable used for Random Forest predictions, `hour` serves as a surrogate here to better visualize the data:

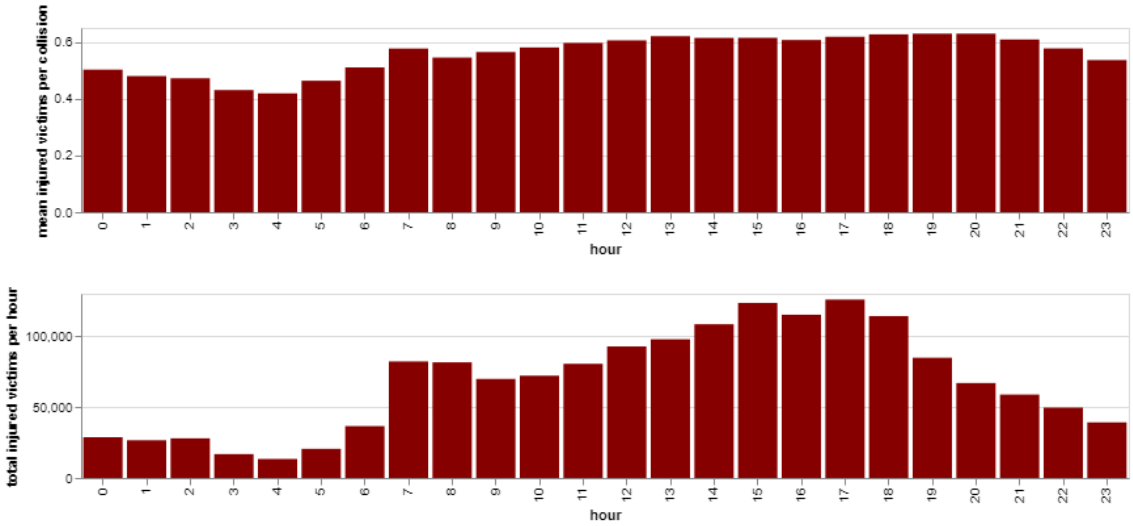

Figure 9: Mean (top) and total (bottom) injuries per hour of the day.

This comparison highlights that the majority of this effect of `hour` on `victims_injured` is driven by the increased number of overall crashes in afternoon/evening hours, but that there is also a difference observed in the average number of victims per crash (maximum 0.63 at 8pm, minimum 0.42 at 4am).

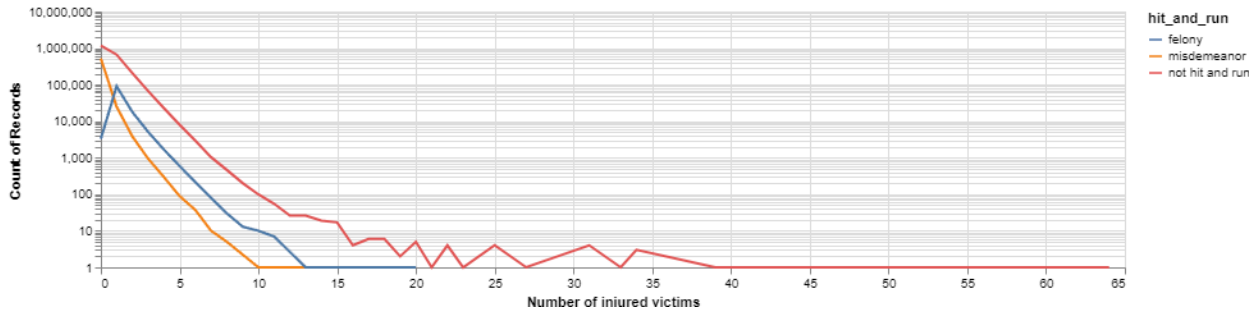

Figure 10: Distribution of the number of injured victims based on `hit_and_run` category

The distribution of injured victims by `hit_and_run` category was also explored and there are a few interesting observations here. The vast majority of collisions are not hit and runs, especially at higher numbers of injured victims. This makes intuitive sense – a crash bad enough to have 20+ injured parties is probably difficult to flee from. Felony hit and runs are more common than misdemeanor hit and runs, except in the cases with only 0 or 1 injuries. This might be explained by crashes without injuries (or with only the driver at fault injured) not as often meeting the threshold for felony status. Felony being more common than misdemeanor otherwise might be explained by drivers being more likely to flee from a higher-consequence major accident than a minor fender bender.

Lasso Regression

Our Lasso models determined that there were 4 significant features in the dataset: `hit_and_run_misdemeanor, party_count, minute, distance`. It is worth noting that these predictors were also ranked highly significant using the Random Forest model so both models reached similar conclusions. The `hit_and_run_misdemeanor` and `minute` were discussed above in the Random Forest discussion section so we will not explore them further. An increase in party count makes intuitive sense as to why this would lead to an increase in the number of injuries. If there are more people involved in a collision then there are more people available to be injured. The distance feature is not as intuitive. This measures the distance of the crash from the intersection with the secondary roadway in feet. This is measuring how far from an intersection a collision took place. High speed accidents that occur on highways would have a large distance from an intersection. This predictor could be suggesting that collisions on large highways result in more injuries.

Conclusion

The findings from our analysis highlight several patterns. We noticed a clear peak late at night for alcohol related accidents, aligning with the time when bars close. On the contrary, non-alcohol related accidents follow typical traffic patterns, with peaks during morning and evening work commute hours.

Our investigation into the number of injuries per collision led to an unexpected result. Our previous assumption was that alcohol involvement would lead to more serious injuries, but we found the opposite to be true, where non-alcohol related accidents were more likely to result in a higher number of injuries. This indicates that other factors are contributing to the severity of collisions.

The time series forecasting for 2020 revealed the impact of the COVID-19 pandemic. The SARIMA model showed discrepancies between the forecasted and actual number of accidents, especially during strict lockdown months. This reflects the role of traffic in the number of collisions, as during this time, there was reduced traffic volume. However, alcohol related accidents didn't experience the same magnitude in decline, indicating that drinking and driving patterns stayed somewhat consistent.

The Random Forest and Lasso models provided additional insight into collision severity. Key predictors include the time of day, the number of parties involved, the hit and run classification, and the distance from intersections. This suggests that certain locations, like highways, may increase the likelihood of severe collisions, and could be worth looking further into by policy makers. Furthermore, the significance seen for the time of day on accidents on many occasions could help with optimizing schedules for first responders.

To conclude, our analysis sheds light on the complex factors that influence traffic collision frequency and severity in Los Angeles County. These insights offer valuable information for policy makers and law enforcement which can be used to try to reduce the number of traffic accidents and optimize responsiveness. Further investigations are encouraged to enhance public safety.