Analysis of California Traffic Collisions

Project Members:

- Mark Lifantsev - marklif@stanford.edu
    - *Question 1*
- Redger Xu - redgerxu@stanford.edu
    - *Question 2*
- Ian Cheung - ian02@stanford.edu
    - *Question 3*
- Nikhil Shanmugham - nikhil07@stanford.edu
    - *Questions 4*

Introduction:

Our project investigates the factors contributing to traffic collisions in California, using a massive (~5,000,000 observations) dataset provided by the California Highway Patrol, and sourced from Kaggle. This dataset spans nearly two decades, from 2001 to 2020, and includes detailed information on collisions, parties involved, and victims, however, we mainly focused on the last year of data. We explored several analytical questions, each focusing on different aspects of traffic collisions, such as the severity of injuries, the impact of Daylight Saving Time (DST) on collisions, and the likelihood of alcohol involvement in collisions given other factors.

We were particularly interested in understanding how various factors like time of day, weather conditions, and vehicle types affect the severity and frequency of collisions. Our motivation was a curiosity to understand collisions, and maybe gain some insights on how to avoid them in our own personal lives, and just to learn something new about public safety and traffic, and maybe even find out how to prevent collisions from a civil engineering standpoint.

Dataset Description:

This data comes from the California Highway Patrol and covers collisions from January 1st, 2001 until mid-December, 2020 with no post-processing. These are full database dumps from the CHP five times, once in 2016, 2017, 2018, 2020, 2021. (according to the publisher/author of the dataset). Each year, there are three separate datasets: collisions, parties involved, and victims involved.

(ii. a&b) Key Features (Observational Units):

***Collisions*:** Includes information on each collision, such as date, time, location, collision type, weather conditions, road conditions, lighting, and collision severity.

***Parties Involved*:** Contains data on all parties involved in each collision, including vehicle types, driver details, age, gender, sobriety, and reasons for the collision.

***Victims Involved*:** Details the victims in each collision, including injury severity, age, gender, seat position, safety equipment used, and other information.

Below we included a link from the author of the dataset which displays the columns with their respective names and possible values in an organized table. The observational units above are only a few of the important ones. A full list is provided in the data dictionary below.

(ii. c) Overall Size:

There are a lot of rows in this dataset (around 5 million). However, due to the big size we may have to reduce the data in this dataset to fit it into colab without running out of the cloud's memory. The attributes are listed in the data dictionary. The missing values are already assigned to a certain value that is listed in the data dictionary. The specific ways we reduced our data are in each question design.

(b) Data Preparation:

We will merge the data by CaseID so we will have access to all the data in one single dataframe and it will give us a holistic view of each traffic collision event. Note: There are specific data preparation steps described in each question.

Data Dictionary:
https://docs.google.com/document/d/1RxeKcv0GZVyPS-__EcUg_JGke3aKGlO2/edit?usp=sharing&ouid=110364097269444718069&rtpof=true&sd=true

Actual Dataset Link:
https://www.kaggle.com/datasets/alexgude/california-traffic-collision-raw-switrs-data

Exploratory Data Analysis:

Question 1:

EDA on Distribution of Severity Levels

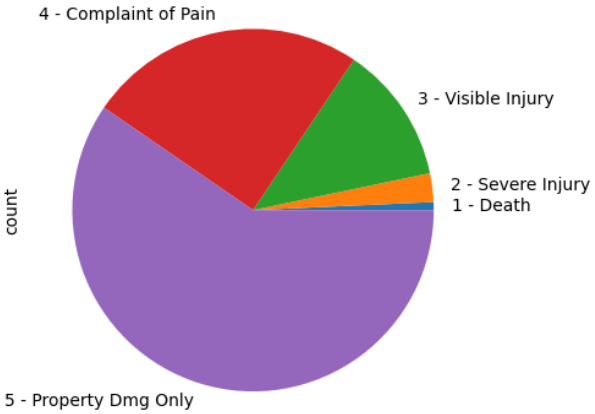


Figure 1.2.1: *Distribution of Severity Levels.*

We can see that we have extremely severe class imbalance. The worst severity levels are extremely underrepresented in the dataset.

EDA on Collision Severity in Relation to *Categorical* Collision Conditions:

For our categorical predictor variables we plotted them against our categorical target variable in the following fashion. Given a predictor feature X and a target feature Y: For each level of X (let us call it x) and for each Level of Y (call it y) calculate the following probability function: $\ln\left(\frac{P(y \mid x)}{P(y)}\right)$. If this value is 0, there is no correlation, and if it is positive, this level of X positively correlates with the level of Y.

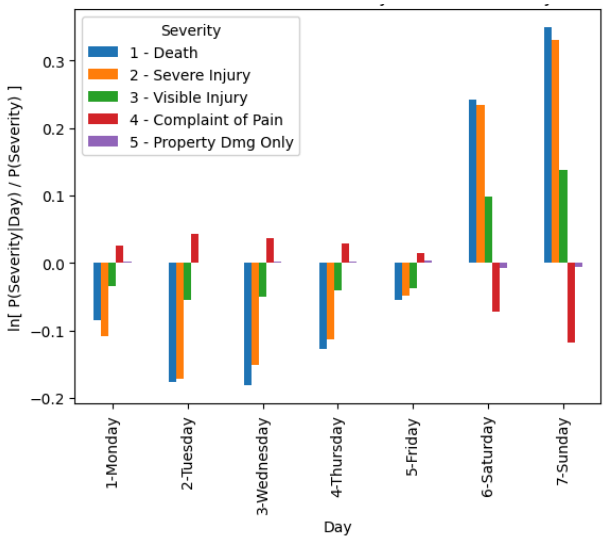
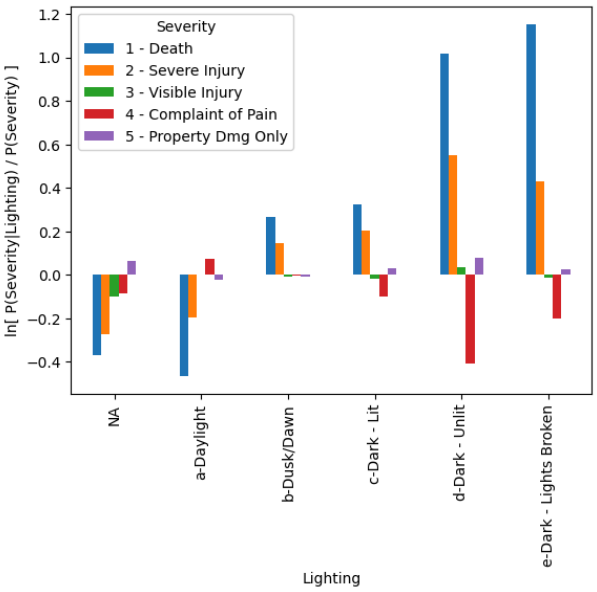


Figure 1.2.2.1: *Correlation between Day of week and Collision Severity*

This plot shows us that on the weekends, a collision is more likely to result in death than on any other day.

Figure 1.2.2.3: *Correlation between Lighting of week and Collision Severity*

Darker streets, quite logically, are a lot more likely to host deadly collisions.



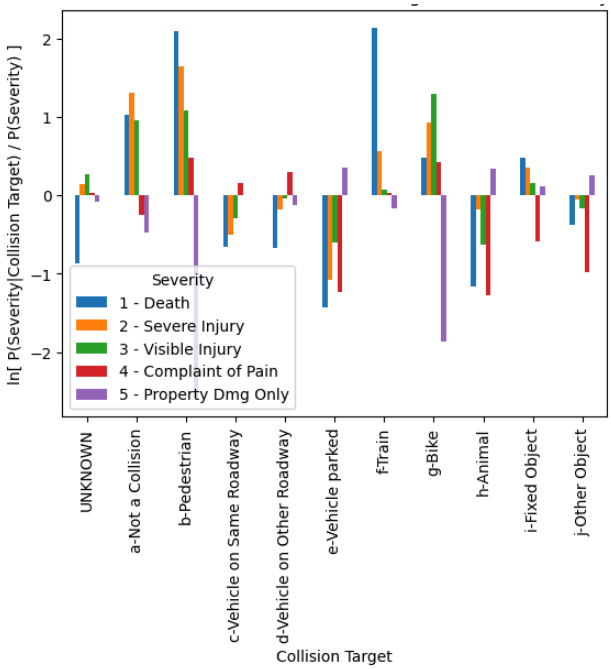


Figure 1.2.2.4: *Correlation between the Object Collided With and Collision Severity*

When someone hits a train (for example if they linger on the tracks), there is a high chance of death. When someone hits a parked vehicle, there is a low chance of death (maybe because you mostly hit parked cars in parking lots ***at low speed****)*

EDA on Collision Severity in Relation to Numerical Collision Conditions:

For our only numeric variable: minute of day, we performed a similar analysis as for the categorical variables. For each minute of the day, we calculated how that affects the chances of death (and the other Severity levels), in a similar fashion, and plotted that on a graph:

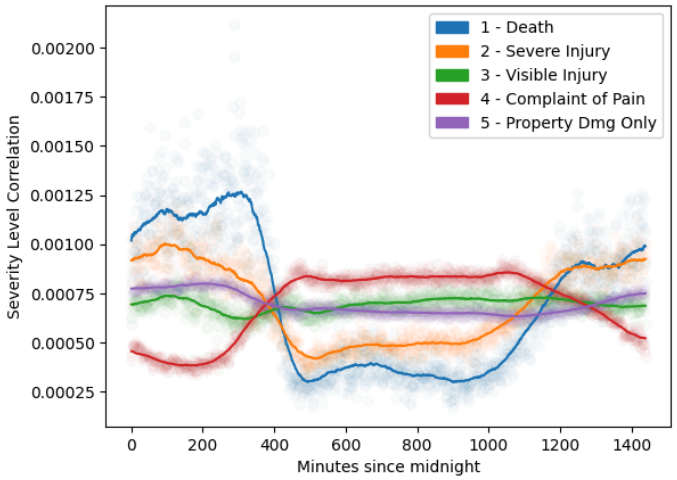


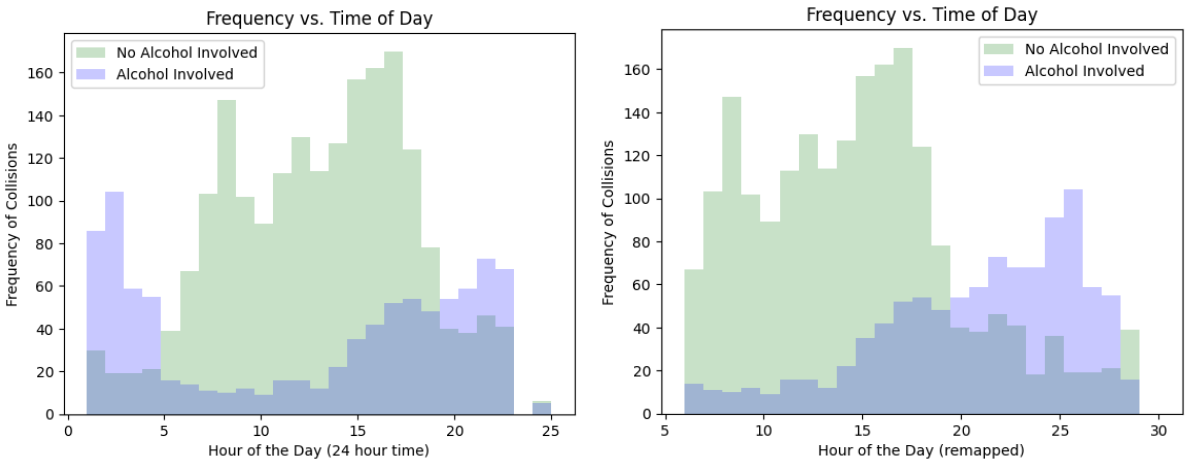
Figure 1.2.3: *Probability of death and other severity levels vs the minute of day.*

The solid line is a heavily smoothed version of the day (moving window average), and the translucent scatterplot is less smoothed, to give an idea of the data variability.

We can see that the chances of death during the day: ~6 AM to ~5 PM are lower than those during the night.

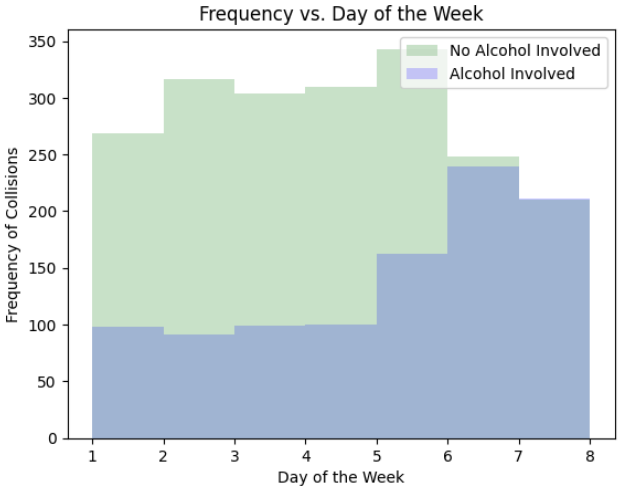
Question 2:

Figure 2.1: *Graph of frequency vs. the hour of the day (selected from training set)*



The graphs above show a histogram of hours of the day vs. the frequency of crashes, with the graph on the right being a remapped version of the time of day. This was done to create a better visualization for the frequency of crashes with alcohol, because it made more sense to have the time go simply from morning to night, rather than from night to day to night.

It was also found that the day of the week had a noticeable relationship with the frequency of alcohol-related collisions, increasing from Monday to Sunday.



Out of all of the numeric variables examined for this question, the hour of the day (mainly the remapped version) and day of the week were found to have the strongest correlations with alcohol involvement.

The categorical variables were not found to have any significant relationships with the involvement of alcohol, but the ones with stronger relationships were used in the second round of model training.

<u>Question 3:</u>

We uncovered the patterns and trends of the frequency of collisions over the past decade. Using that, we were able to predict the "future" collision frequencies.

We also uncovered the most accident prone location: the area around California State Route 60 and Grand Avenue in Diamond Bar, LA.

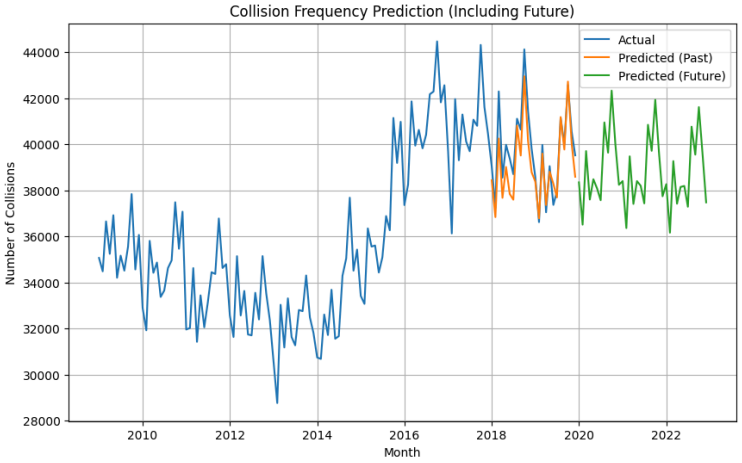


Figure 3.6 *- Collision Frequency Prediction*

YEAR	PRIMARY_RD	SECONDARY_RD	COUNT
2009	RT 60	GRAND AV	325
2010	RT 60	GRAND AV	349
2011	RT 60	GRAND AV	353
2012	RT 60	GRAND AV	343
2013	RT 60	GRAND AV	377

2014	RT 60	GRAND AV	371
2015	RT 15	RT 138	230
2016	SR-60 W/B (POMONA FREEWAY)	GRAND AVE	283
2017	SR-60 W/B (POMONA FWY)	GRAND AVE	202
2018	SR-60 W/B (POMONA FWY)	GRAND AVE	205
2019	SR-60 W/B (POMONA FWY)	GRAND AVE	223

Table 3.9 *- Top Accident-Prone Location by Year*

** This is just a preview. Full details are included in the* ***Results*** *section of Question 3.*

Questions 4:

One of the initial steps for these two questions was to visualize the distribution of traffic collisions over time. We plotted the number of collisions on a monthly basis and created histograms to examine how collision frequency varied across different periods. This allowed us to observe fluctuations in collision frequency and better understand the temporal patterns in the data.

We also investigated the severity of collisions across the dataset. By generating count plots to visualize the distribution of collision severity, we were able to see how often different severity levels occurred.

Additionally, we examined the impact of weather conditions and road surface states on traffic collisions. We visualized the distribution of collisions under various weather conditions and road surface states and identified times when adverse weather and poor road conditions were more common. This analysis emphasized the need to include weather and road surface conditions as features, as they are likely to influence collision risk.

In terms of data quality, we conducted detailed cleaning. We performed value counts on the features, which helped us identify and address instances of random or incorrectly formatted data. We matched our data to the official SWITRS data dictionary, replacing or correcting values as needed. This included standardizing unknown or out-of-range values to "UNKNOWN" to maintain consistency across the dataset.

Question 1 - Notebook

1. **Analytical Question**

   "*Predict the severity of injuries in traffic collisions based on factors like time of day, vehicle type, and weather conditions.*"

   The goal is, given the conditions of a collision, be able to predict a general idea of if the collision is deadly, or if they will walk away with a bit of pain.

2. **Design**

   Instead of trying to quantify the injury severity in a more descriptive way, we will be predicting the categorical variable `COLLISION_SEVERITY`, which has five levels and describes the severity of the worst injury suffered in the collision. This is a classification problem with mostly categorical predictor variables.

   All data is taken from the California Traffic Collision Raw SWITRS Dataset. These are the features:

   a. Daytime:
      i. `COLLISION_TIME`
         1. **Note:** in the original dataset, this variable is in HHMM format, we will have to preprocess it to convert into minutes since midnight (so that it is continuous: no jump from 1959 to 2000)
      ii. `DAY_OF_WEEK`
      iii. `LIGHTING`
   b. Conditions:
      i. `WEATHER_1 & WEATHER_2`
      ii. `ROAD_SURFACE, ROAD_COND_1`
   c. Vehicles Involved:
      i. Motor Vehicle Involved With (`MVIW`)
      ii. Statewide Vehicle Type At Fault (`STWD_VEHTYPE_AT_FAULT`)

We only need to take data from collisionrecords.txt so no merging of dataframes is necessary.

For each of the following models we will use a 10 fold grid search to find the best hyperparameters, and then compare the accuracy of our best models to choose the final result.

a. Model 1: KNN Classifier with `k`: `[1, 200]`
b. Model 2: Support Vector Machine Classification
c. Model 3: Logistic Regression

We will use `f1-macro` as our main tool to evaluate the models we produce. In order to gain a deeper understanding of what our classifiers are doing we will examine and interpret the confusion matrices of our top models.

3. Implementation

Note that the dataset has around 5,000,000 observations. When training our models, we cannot possibly train it on the entire dataset, so we need to sample smaller sections of the dataset. In the EDA, we found that the class imbalance is really bad. So in the data sampling, we should try to minimize the imbalance within our sample.

What we did was when we take, for example, a sample with 10,000 observations, we randomly sample 2,000 (10,000 / 5 severity levels) observations of severity 1. Then 2,000 observations of severity 2, and so on. If there are not enough observations of a certain severity level to meet this quota, we just allow there to be some class imbalance (which is still a lot less than without this method)

4. Results

Final Model Parameters: After running all of our grid searches, these are the best parameters for our 3 models.

KNN	SVM	Logistic

`n_neighbors = 36`
`p = 1 (manhattan)`
	`C = 2.282`
	`gamma = auto`
	`kernel = rbf`
		`C = 1.216`
		`penalty = l1`
		`solver = saga`

Figure 1.4.1: *Final Model Performances (run on full dataset: ~5mil observations)*

Model:	KNN	SVM	Logistic
Accuracy	0.3134	0.2645	0.3768
Lenient Accuracy	0.7604	0.7535	0.7922
F1 Score	0.3485	0.2608	0.4132

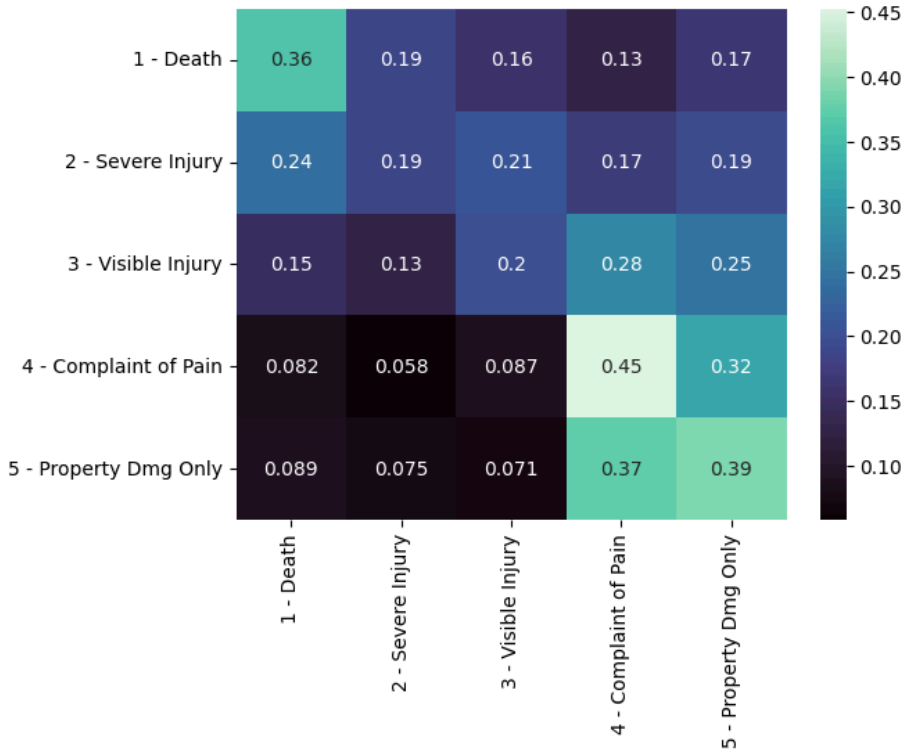
As we can see, the logistic regression model is better than the other two in all of our metrics.

Note: Lenient Accuracy

For this classification, our different categories are not necessarily extremely distinct. For example, severe injury and visible injury are pretty similar. So we should maybe use a metric that, for an observation whose severity = death, would penalize a prediction of 'property damage only' more than a prediction of 'severe injury'.

So, we created a metric, lenient accuracy, that for an observation that has a severity of 'death', counts a prediction of 'death' as 1 correct prediction, a prediction of 'severe injury' as 0.8 correct predictions, a prediction of 'visible injury' as 0.6 correct predictions, etc. It sums these up and divides by the total amount of predictions to report the final 'lenient accuracy'

Figure 1.4: *Confusion Matrix for Best Model: Logistic Regression*



We can see that the diagonal has the highest values: our model is making accurate predictions the majority of the time. However, the cells adjacent to the diagonal also have pretty high values. This is because a collision that results in property damage only is similar to one that results in only a complaint of pain. This is the exact reason we created the lenient accuracy metric.

Overall, this model does not make any outrageous predictions, the cells completely off of the diagonal have low values. However, the accuracy, in the grand scheme of things, is admittedly not ideal. Our speculation on why this might be is that even if the conditions of the collision are exactly the same. A 100 year old passenger might die while a 20 year old might just get injured. These models did not have access to this kind of passenger info, which may explain why our confusion matrix's diagonal is so 'blurry': a lot of predictions slightly off of the diagonal.

Question 2 - Notebook

1. **Analytical Question**

   "*Classify and predict the likelihood of DUI-related collisions based on temporal and environmental factors.*"

   This question aims to examine the factors that contribute most to DUI-related accidents in California. The interest in this question comes from the question of whether or not there are patterns involved in drunk driving, or if the likelihood of alcohol being involved in a collision is uniform. The results of this question aim to help provide more insight to when drunk driving is more likely to occur, improving public safety by allowing highway patrol to stop more potential crashes during high times for alcohol-related collisions.

2. **Design**

   For this question, we utilized the California Traffic Collision Raw SWITRS Dataset, which is a dump of collision-related records from the California Highway Patrol, specifically the 2021 June 4th dump. For this question, the collision time, alcohol involvement, time of collision, date of collision (as a calendar date), day of the week, type of weather, type of location (highway, intersection, ramp), and type of ramp intersection (ramp exit, mid ramp, ramp entry, etc.).

   To prepare the data, in addition to filling in missing/removing missing values, the month, day of month, year, hour, and minute were created as separate columns from the aforementioned variables. A separate measurement of hours, mapping times 12:00 AM to 6:00 AM to hours 24-30 was also created.

   To solve the question, three models were created: a KNN classifier, logistic regression, and SVM. These were put through GridSearchCV to select the best set of hyperparameters (further discussed under Implementation), and

for each grid search, a new model was made with the best hyperparameters from the grid search that only took a selection of the original features as input (the hour remapping, the day of week, the secondary weather measurement, location type, and type of ramp intersection).

To evaluate each model, we used accuracy, precision, recall, and F1 score, but the ultimate deciding factor was F1 score. This was because the data set was imbalanced, so this would mitigate the problem of the models choosing the same class every time.

3. Implementation

To train the model, a randomly generated selection of 2000 non-alcohol collisions and 1000 alcohol involved collisions were combined to create a training set. This was due to performance reasons as well as trying to offset the severe imbalance that the full data set had, while still preserving some element of imbalance to optimize F1 score.

Each of the models used in this question utilized one-hot encoding for categorical features and standardization for numerical features. The standardization was implemented in order to fix an error with the fitting of one of the models, as well as to ensure that all variables could have an equal contribution regardless of scale. In addition, the alcohol involvement variable was converted to numeric values (N to 0 and Y to 1), which was stored in the `ALC` column in the code.

For hyperparameter tuning, the KNN model used K values from 2 to 40 and both Manhattan and Euclidean distance. For the logistic regression, C values for every 0.2 between 0.5 and 2.9 were used (i.e. 0.5, 0.7, ..., 2.7, 2.9), and the solvers used were "`liblinear`" and "`saga`". For the SVM model, the same values as the logistic regression were used for C (0.5, 0.7, ... 2.9), the "`linear`", "`poly`", and "`sigmoid`" kernels were used, and the polynomial degree ranged from 2 to 4.

4. Results

The following results were evaluated on the first 10000 rows of data after cleaning (rows with NaN values were dropped)

Model	Accuracy Score	Precision	Recall	F1 Score
KNN with Euclidean distance, `K=19`	0.845	0.657	0.307	0.419
Logistic Regression with `C=2.9, penalty=l1, solver=saga`	0.863	0.577	0.326	0.417
SVM with `C=2.3, degree=2, kernel=poly`	0.861	0.607	0.327	0.425

Table 2.4.1: *Models with best hyperparameters, using all features*

Model	Accuracy Score	Precision	Recall	F1 Score
KNN with Euclidean distance, `K=19`	0.836	0.659	0.292	0.405
Logistic Regression with `C=2.9, penalty=l1, solver=saga`	0.862	0.581	0.325	0.417
SVM with `C=2.3, degree=2, kernel=poly`	0.873	0.531	0.341	0.415

Table 2.4.2: *Above models trained on a subset of the original features*

The models all had decent accuracy on the testing data, however this is not extremely significant given the imbalance of the testing set. For the other three metrics, the models were not very effective, but the first SVM model (as shown in 2.4.1) trained on all of the features used in this question had the best F1 score. The SVM model trained on the subset of features (as shown in 2.4.2) had a slightly lower F1 score, but the highest accuracy of all six models. Ultimately, the first SVM model is the best model based on the evaluation criteria (F1 score).

Question 3 - Notebook

1. **Analytical Question**

   "*Compare how the frequency and type of traffic collisions changed in California and the most-prone collision locations over time?*"

   This question aims to analyze how often traffic collisions occur in California and the types of collisions that happen most frequently, and how these patterns have changed over the years, both in California overall and in locations known for having a lot of accidents. This is interesting because it helps us understand if driving habits, road conditions, or safety measures are making a difference in reducing accidents. It's relevant for improving road safety and preventing future collisions.

2. **Design**

   For this analysis, we are using a dataset of traffic collision records in California from 2009 to 2019. The dataset includes details like the date and time of each collision, the type of collision, the road conditions, and the location. We chose this time frame in this question to avoid inconsistencies in 2020 data due to the COVID-19 pandemic.

   Data Preparation:

   1. We extracted year, month, date, time, and day of the week from the columns `COLLISION_DATE, COLLISION_TIME, DAY_OF_WEEK` to analyze trends over different time periods by creating a new column called `COLLISION_DATETIME` with datetime objects.
   2. We mapped day numbers to day names for better readability.
   3. We filtered the data to focus on the years between 2009 and 2019.

   Methods:

   1. We used visualizations like bar charts and line graphs to illustrate the trends in collision frequency and types over time.

2. We grouped the data by year, month, and collision type to calculate the frequency of collisions and the percentage of each collision type.
3. We used the `skforecast` library to predict future collision counts based on historical trends, evaluating and perfecting the model using RMSE.
4. We identified the top collision locations by grouping collisions by primary and secondary roads and counting the number of collisions at each location, as well as analyzing the trend.

3. Implementation

To handle the large dataset efficiently, we leveraged Dask DataFrames instead of Pandas. We read the CSV file into a Dask DataFrame using `dd.read_csv()`, and only import certain columns using the usecols attribute. We also handled missing values by setting the `na_values` attribute. Only after the preprocessing of data did we convert the Dask DataFrame back to Panda DataFrame using `.compute()`.

In addition, to facilitate the analysis of trends over different time periods, we created a temporary column called `YearMonth`. This column combines the year and month information from the `COLLISION_DATE` column, allowing us to easily group and analyze collision patterns by month and year.

Moreover, we investigated the most common PCF violation categories and collision types for the top locations in each year to gain extra insights into the factors contributing to collisions at these locations.

For visualizations, we used bar charts to visualize the frequency of collisions by date, month, year, and type; and we used line graphs to illustrate the percentage of collision types over time and the trend in collision frequency for the top locations.

4. Results

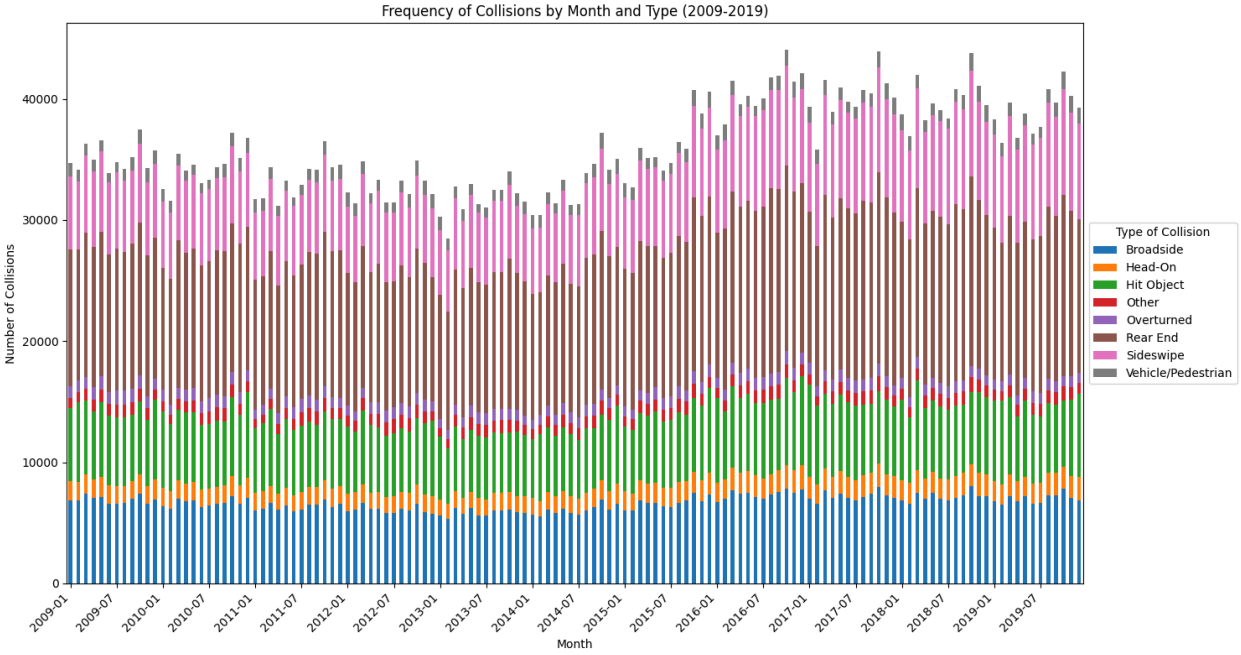


Figure 3.1 - *Frequency of Collisions by Month and Type* (2009-2019)

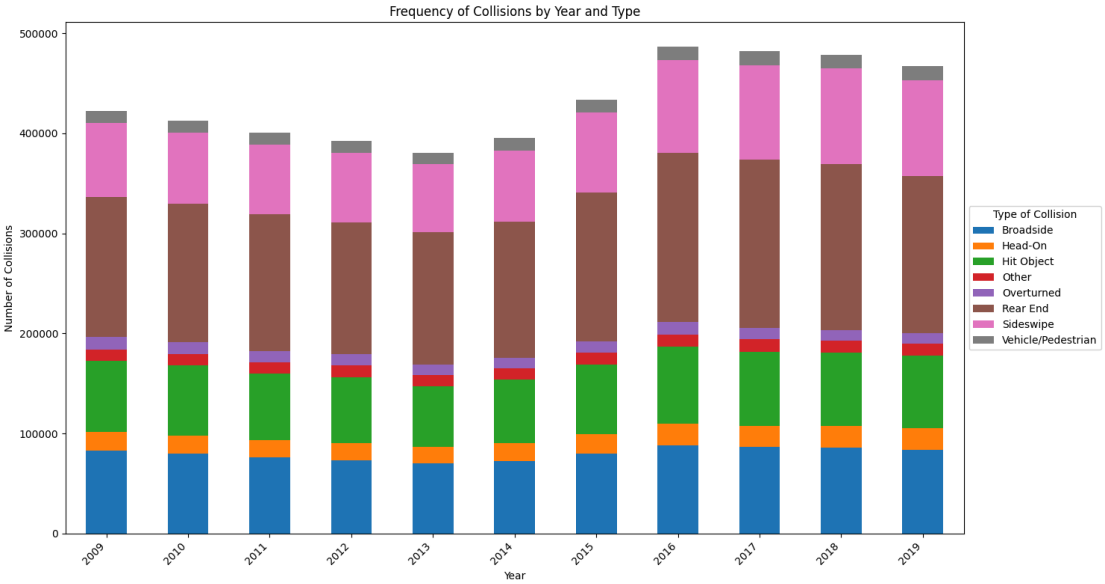


Figure 3.2 - *Frequency of Collisions by Year and Type* (2009-2019)

We can notice that the trend of collision frequencies has been upwards throughout the years of 2009 and 2019. From Figure 3,2, specifically, it decreased from 2009 to 2013, but rose relatively dramatically from 2013 to 2016, and then gradually descended from 2016 onwards.

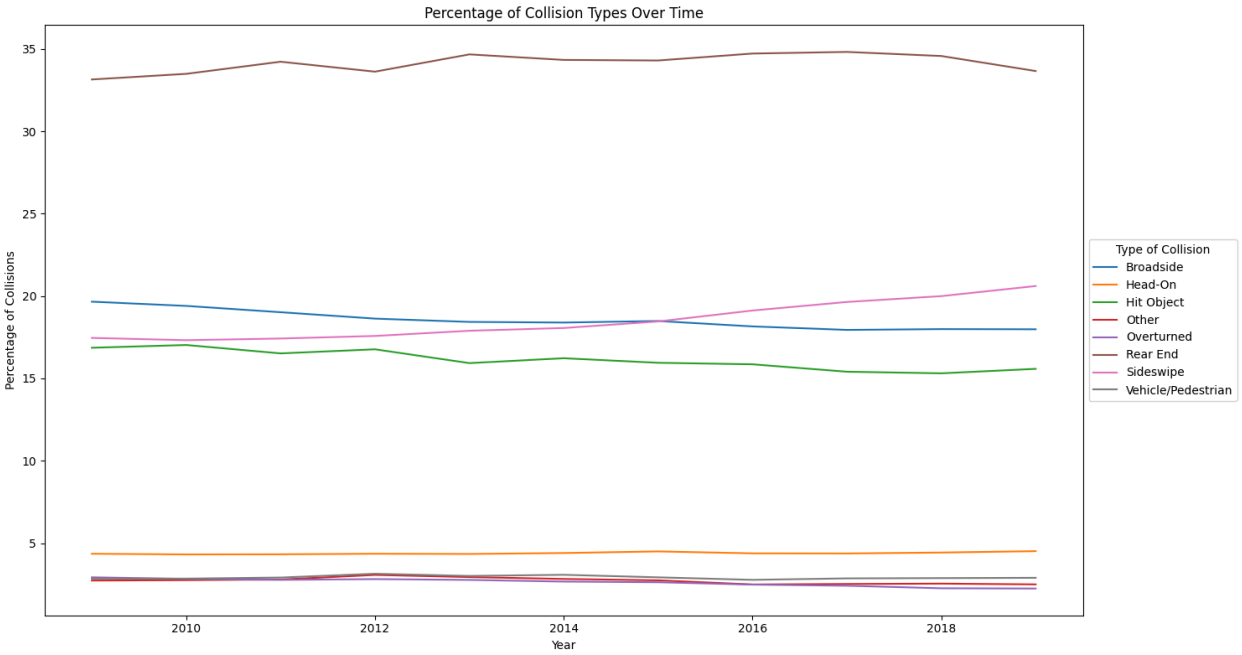


Figure 3.3 - *Percentage of Collision Types Over Time*

It is hard to analyze the percentage of collision types using the stacked bar graphs above, so Figure 3.3 can help clarify a little bit. We can see that "Sideswipes" are increasingly becoming more common among the types of collisions, though only reaching around 22 percent. The majority of collisions are "Rear Ends".

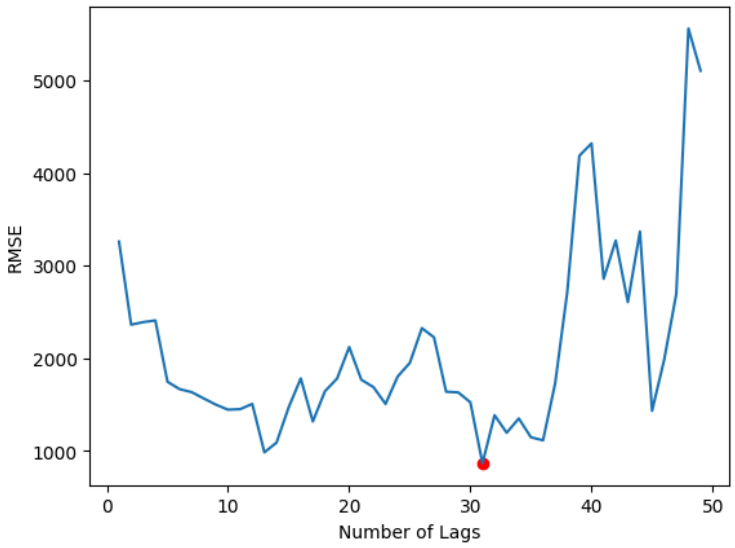


Figure 3.4 - *RMSE vs Number of Lags*

We then try to predict the collision frequencies from 2020 onwards. (As *mentioned earlier, we analyze our data up to 2019 as COVID-19 led to significant decrease in collisions due to stay-at-home policies. In a sense, we are predicting the collision frequencies if COVID-19 did not happen).*

Since we are using Time Series Forecasting, we want to **find the best "lag" value**. In particular, the value of the lag denotes the past time interval we want to use to predict the future. From Figure 3.4, we can see **31 lags returns the lowest RMSE** (863.5), indicating that using the prior 31 months is the best model. *Please also note that we used the most recent 24 months as testing (but these 2 known years are still included in the training model, as the objective is to predict unknown future data).* Below in Figure 3.5, you can see the comparison of our model prediction and the actual data, side-by-side.

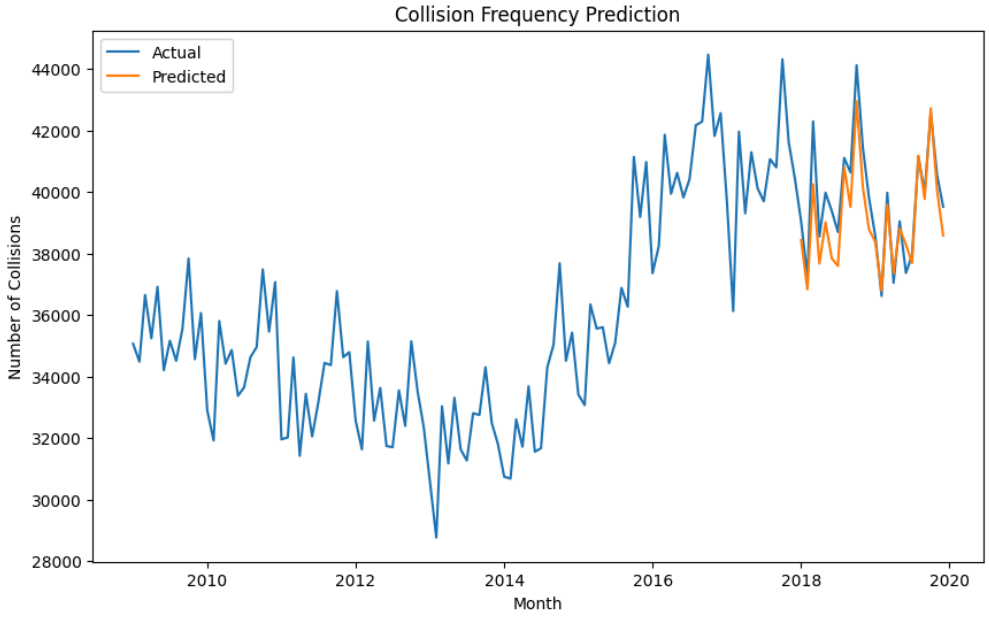


Figure 3.5 *- Collision Frequency Prediction of Most 2 Recent Months* (2018-2019)

Now it is time to predict the future, let's say for the next 3 years!

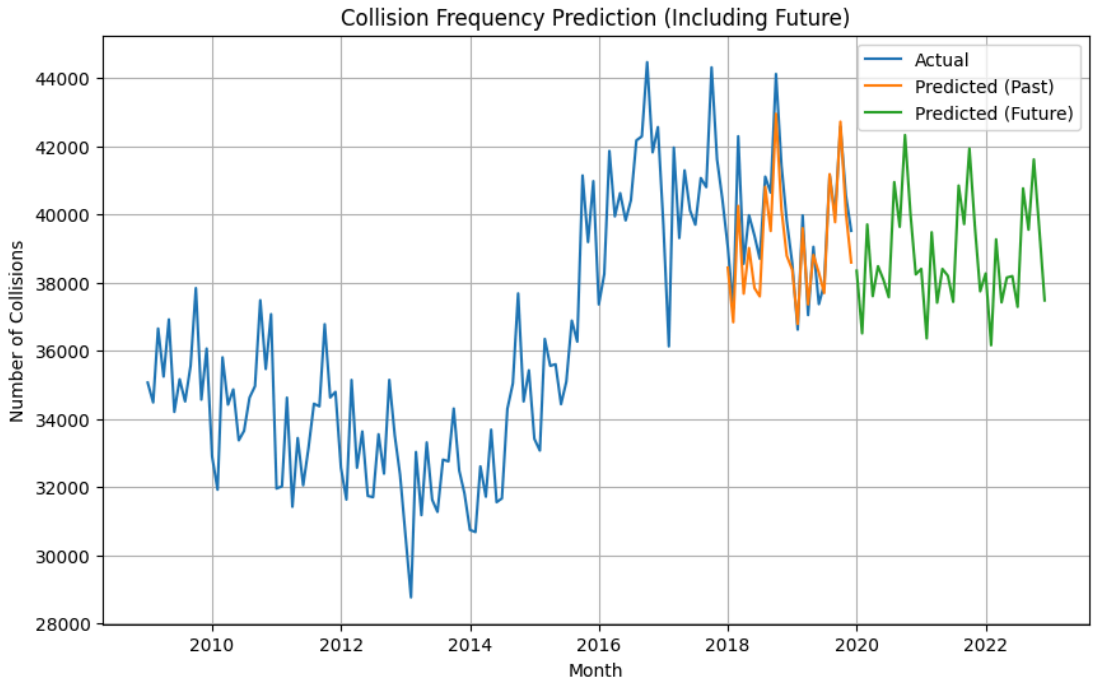


Figure 3.6 - *Collision Frequency Prediction of Future 3 Years (2020-2022)*

It seems like, throughout the entire graph in Figure 3.6, that there is a clear shape of how the collision frequencies vary over each year. Specifically, t**he start of the year usually starts with a lower frequency**, while **the frequency peaks at around November**. What an interesting pattern!!!

Now let's switch gears for a little bit.

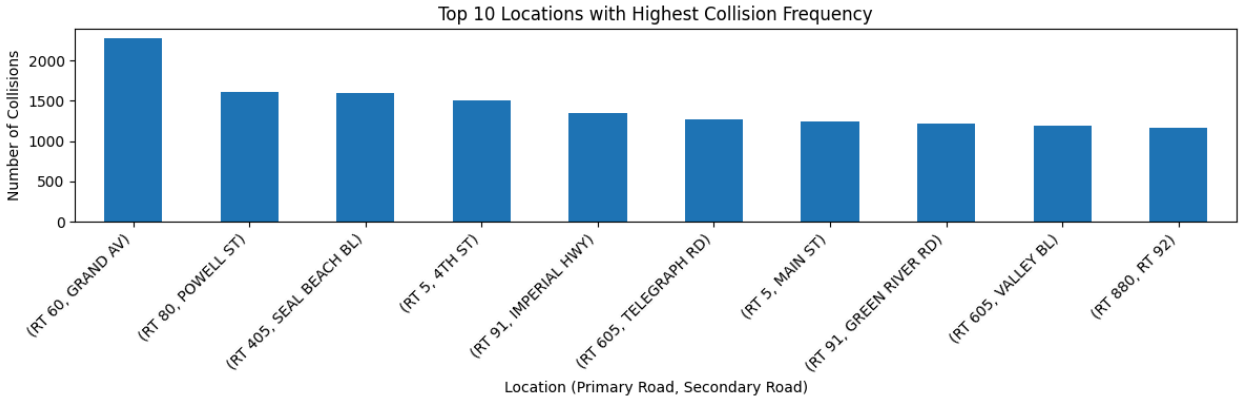


Figure 3.7 - *Top 10 Most Collision Prone Locations*

From Figure 3.7, it is clear that the **intersection around California State Route 60 and Grand Avenue** has a whopping high number of collisions throughout the entire dataset. Let's plot the collision frequency trend for each of the 10 most collision prone locations, as shown below in Figure 3.8.

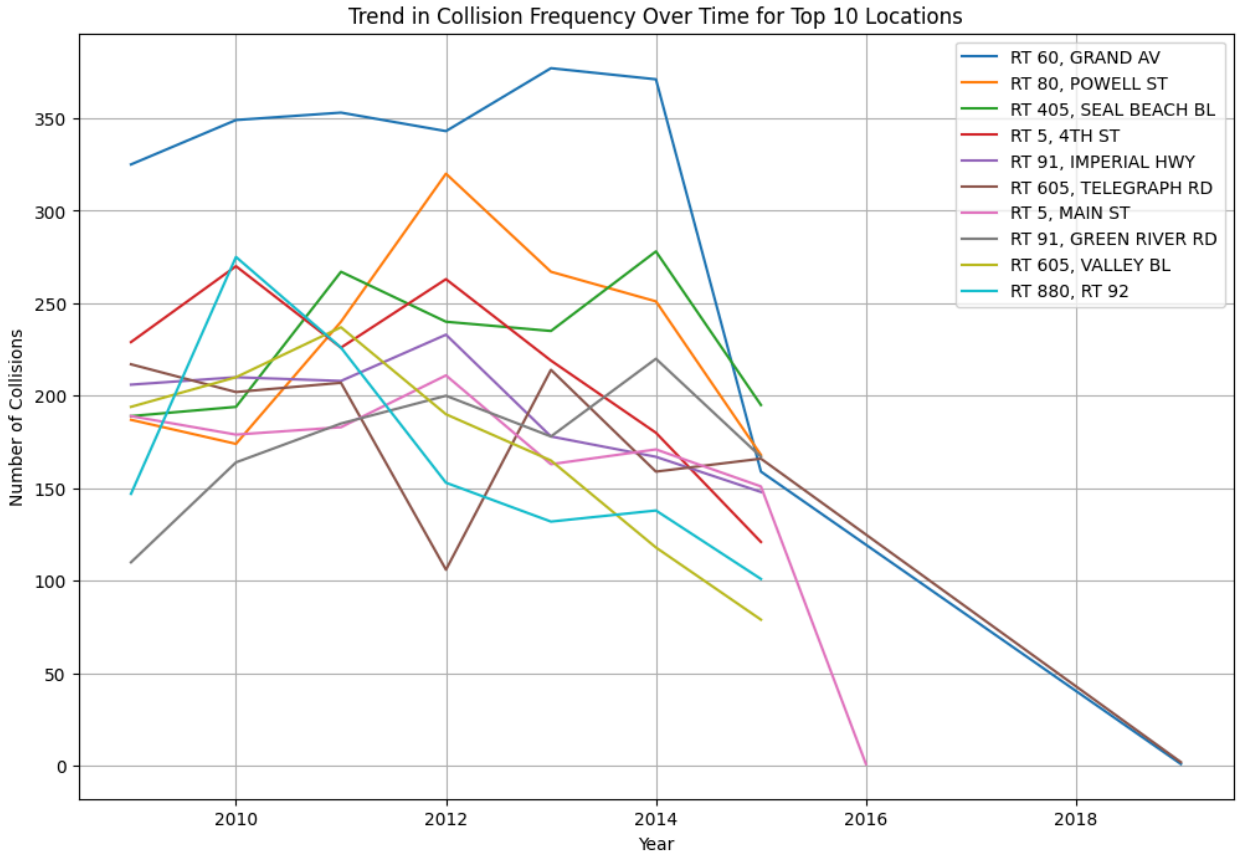


Figure 3.8 *- Trend in Collision Frequency Over Time for Top 10 Locations*

The right side of the graph looks off! There is no way that there are suddenly no collisions in these locations. Let's see what happens if we try to find the top location of collisions, rather than overall, but instead **by year.**

YEAR	PRIMARY_RD	SECONDARY_RD	COUNT
2009	RT 60	GRAND AV	325
2010	RT 60	GRAND AV	349
2011	RT 60	GRAND AV	353

2012	RT 60	GRAND AV	343
2013	RT 60	GRAND AV	377
2014	RT 60	GRAND AV	371
2015	RT 15	RT 138	230
2016	SR-60 W/B (POMONA FREEWAY)	GRAND AVE	283
2017	SR-60 W/B (POMONA FWY)	GRAND AVE	202
2018	SR-60 W/B (POMONA FWY)	GRAND AVE	205
2019	SR-60 W/B (POMONA FWY)	GRAND AVE	223

Table 3.9 *- Top Accident-Prone Location by Year*

Now we realize that the format of the data has simply changed (phew!). Specifically, the location `(RT 60, GRAND AV)` has changed to `(SR-60 W/B (POMONA FWY), GRAND AVE)`. However, we can see that the most accident-prone location (except 2015) is still that same intersection around California State Route 60 and Grand Avenue.

We found the location on Google Maps, and it is suspected that the collisions are most likely resulting from the **poor highway design**, where two highways have to completely merge together before splitting again. To further understand this phenomenon, we found that the top PCF Violation and collision type among all the above collisions are “Unsafe Speed” and “Rear End” respectively.

As discovered in an recent article about a road change in the area, the location is indicated as the **‘worst truck bottleneck’ in the nation**, where congestion (hence, collisions!) frequently occurs. This suggests that the Federal Highway Administration should improve the road design of the surrounding areas around California State Route 60 and its exit of Grand Avenue.

Question 4 - Notebook

1. **Analytical Question**

   "How do the start and end of Daylight Saving Time (DST) affect the likelihood of traffic collisions in California, and can we classify these periods based on collision data (post-DST start, post-DST end, normal time)?"

   This question aims to explore the impact of the start and end of Daylight Saving Time (DST) on the likelihood of traffic collisions in California. The interest lies in understanding whether the time shifts associated with DST—specifically, the loss or gain of an hour—affect driving behavior, potentially leading to an increased number of collisions. This analysis is particularly important for public safety and policy-making, as it could inform decisions on whether DST should be modified or maintained, based on its effects on road safety. The goal is to classify traffic collisions into three categories: post-DST start, post-DST end, and normal time, to identify if there are distinguishable patterns in collision data that correlate with these periods.

2. **Design**

   To address this question, we utilized the California Traffic Collision Raw SWITRS Dataset, which includes detailed records of traffic collisions across the state. The key variables selected for analysis were Collision Date, Collision Time, Vehicle Type, Road Surface Conditions, Weather Conditions, Lighting, and Control Device.

   Data Preparation

   1. A new categorical variable, `DST_PERIOD`, was engineered to categorize each collision into one of three classes: post-DST start, post-DST end, or normal time. This was based on the Collision Date.
   2. The dataset was initially found to be highly imbalanced, with a significantly larger number of collisions classified as occurring during the "normal" time period. To ensure a more balanced and effective

analysis, the data was reduced, particularly by undersampling the "normal" category, to create a more even distribution across all three DST periods.
3. The target variable was `DST_PERIOD`, and the features included both categorical and numerical variables related to the conditions surrounding each collision.

Three models were employed:

1. Logistic Regression: A baseline model was developed using Logistic Regression to classify collisions into the three DST periods. The OneHotEncoder and StandardScaler transformations were applied to the features through a column transformer. The Logistic Regression model was then integrated into a pipeline, which handled both preprocessing and model training in a single flow.
2. Random Forest Classifier: Captures potential non-linear relationships and interactions between the features. It is similar to the Logistic Regression model because the Random Forest Classifier was also encapsulated within a pipeline with the same preprocessing steps.
3. Support Vector Machine (SVM): The SVM model was used to explore the separability of the DST periods in a high-dimensional feature space. An SVM pipeline was created with the `rbf` kernel to manage complex decision boundaries between the classes.

Each model was incorporated into a pipeline, which included preprocessing steps like One-Hot Encoding for categorical features and Standard Scaling for numerical features. Hyperparameter tuning was conducted using `GridSearchCV` with a 5-fold cross-validation.

3. Implementation

The implementation phase involved executing the technical design through the following steps:

Preprocessing Pipelines:

- Categorical Features: OneHotEncoder was applied to categorical variables (`COLLISION_SEVERITY`, `WEATHER_1`, `ROAD_SURFACE`,

LIGHTING, and CONTROL_DEVICE) to convert them into a format suitable for model input.
- Numerical Features: StandardScaler was used to normalize HOUR and DAY_OF_WEEK, ensuring that all features contributed equally to the model.

Model Development:
- Logistic Regression Pipeline: This pipeline integrated the preprocessing steps with the Logistic Regression model. The ovr (one-vs-rest) strategy was applied to handle the multi-class classification problem.
- Random Forest Pipeline: Included the same preprocessing steps, followed by the Random Forest model to capture complex interactions between features.
- SVM Pipeline: Used the rbf kernel to manage non-linear decision boundaries, critical for accurately classifying the DST periods.

Hyperparameter Tuning:
- Logistic Regression: Parameters such as C, penalty, and solver were tuned.
- Random Forest: Parameters like n_estimators, max_depth, and min_samples_split were optimized.
- SVM: The C and kernel parameters were adjusted for optimal performance.

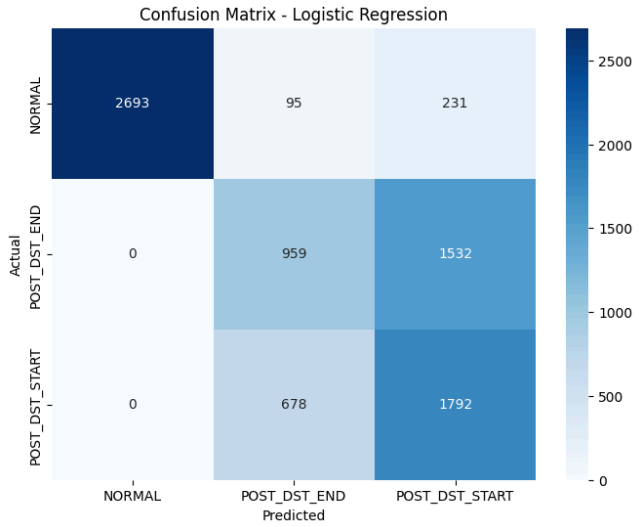
Evaluation:
- Each model's performance was evaluated on a holdout test set using accuracy, precision, recall, and F1-score. Confusion matrices were generated to provide a visual breakdown of the models' predictions.

4. **Results:**

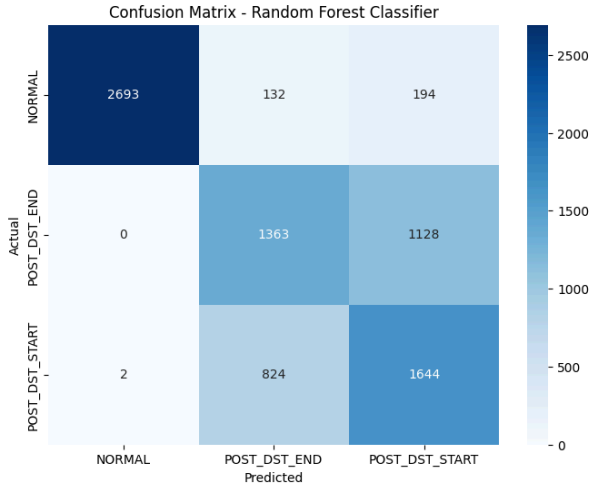
The results from the three models are summarized below:

1. Logistic Regression:
    - Best Parameters: C=0.01, penalty='l1', solver='saga'
    - Best Cross-Validation Accuracy: 0.6853
    - Test Accuracy: 0.6822

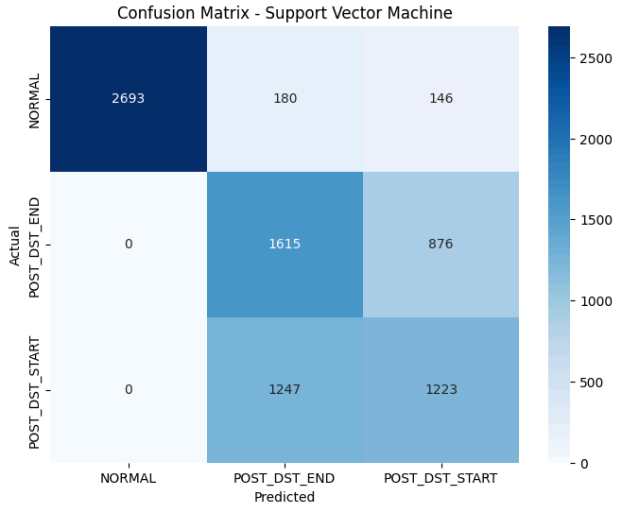
- Precision: `0.7072`
- Recall: `0.6822`
- F1-Score: `0.6826`
- Confusion Matrix:



2. Random Forest Classifier:
    - Best Parameters: `n_estimators=200, max_depth=15, min_samples_split=5`
    - Best Cross-Validation Accuracy: `0.7147`
    - Test Accuracy: `0.7143`
    - Precision: `0.7331`
    - Recall: `0.7143`
    - F1-Score: `0.7207`
    - Confusion Matrix:



3. Support Vector Machine (SVM):
    - Best Parameters: `C=1, kernel='rbf'`
    - Best Cross-Validation Accuracy: `0.6984`
    - Test Accuracy: `0.6931`
    - Precision: `0.7127`
    - Recall: `0.6931`
    - F1-Score: `0.6995`
    - Confusion Matrix:



The models indicate a higher likelihood of collisions in Post-DST Start, which aligns with the hypothesis that the shift in time, particularly the loss of an hour, may lead to increased driver fatigue and changes in morning light conditions. These factors could contribute to a higher incidence of collisions. The confusion matrices provide a visual breakdown of how well each model performed across the different DST periods, with the Random Forest model achieving the highest accuracy and F1-score. This model shows a better balance between precision and recall, especially for the `POST_DST_END` and `POST_DST_START` categories.

Discussion and Conclusions

The common theme across our project was analyzing and predicting traffic collision dynamics in California based on various temporal, environmental, and human factors. Each question we explored on a different aspect, such as the role of alcohol and Daylight Saving Time or the influence of weather and road conditions on collision severity, and more.

We were able to gather many insights from the data. For instance, we found that the likelihood of alcohol-related collisions was significantly influenced by the time of day and day of the week, with higher frequencies observed during late hours and weekends. Similarly, the start and end of Daylight Saving Time were associated with noticeable shifts in collision patterns, suggesting that the time change may impact driver behavior and collision risk.

Moreover, the analysis of collision severity revealed the complexity of predicting future outcomes based on past data. While our models achieved moderate accuracy, the difficulty in distinguishing between different severity levels highlights the need for better data analyzing and modeling techniques. Nevertheless, the Random Forest and SVM models usually performed better in capturing non-linear relationships between features, but there is still room for improvement.

All in all, our project demonstrates the potential of understanding past traffic collisions and informing public safety initiatives through the power of data science. For example, we can inform targeted interventions and appropriate resource allocation by local and state authorities. Ultimately, it would hopefully lead to the implementation of more effective safety measures, ultimately reducing collision rates and saving lives.