INDENG 142 Final Report:
Predicting Search of Persons and Properties From Traffic Stops in San Francisco

Anishi Patel, Brianna Cortes, Soumya Agarwal, Nhu Vu, Shobhan Mangla, Aditya Pattani

Motivation

In light of prevalent concerns regarding racial biases and profiling within law enforcement, our project aims to conduct a comprehensive analysis using the San Francisco Police Department (SFPD) Stop Data dataset. This dataset, established to comply with the Racial and Identity Profiling Act (RIPA) or California Assembly Bill (AB) 953, captures detailed information on police stops, including circumstances and perceived identity characteristics of individuals stopped by SFPD officers.

Our primary goal is to identify and quantify existing biases within these stop data records. We intend to explore how factors such as age, race, ethnicity, english proficiency, and gender influence the likelihood of being subjected to bodily and property searches during police stops conducted by the SFPD. The decision to focus on this dataset is informed by personal experiences within our team—which includes multiple international students who have encountered unwarranted stops by US authorities, mirroring broader patterns of racial profiling.

By leveraging this data spanning from July 1st, 2018, to the present, we aim to shed light on disparities and correlations within SFPD stop practices. San Francisco, renowned for its diversity, serves as an ideal setting for our analysis. As UC Berkeley students, our findings will not only offer insights into the dynamics of our immediate community but also contribute to a broader discourse on the impact of biased law enforcement practices in culturally rich urban environments.

Through this project, we aspire to contribute to the ongoing dialogue on police accountability and advocate for reforms that promote fairness and transparency in law enforcement interactions, particularly concerning marginalized communities.

Data Source

We obtained our dataset from the official DataSF website, specifically the dataset titled "Count of Individuals Stopped by Reason for Stop by month, July 2018 to latest," which was listed under the Public Safety sector. The original, raw dataset consisted of 85 features and 250,093 rows, and even though the dataset is continually updated to reflect traffic stops on the daily, we are only using data up until the point of us acquiring the dataset to simplify our task.

Furthermore, there are various features that describe relevant information about each traffic stop that occurred in San Francisco and the larger Bay Area. These features include but is not limited to the following: stop_datetime, duration_of_stop, location, city, is_lgbt, perceived_race_ethnicity, perceived_gender, perceived_age, reason_for_stop, reason_for_stop_code, perceived_or_known_disability, actions_taken, and actions_taken_code. That being said, because of the sheer size of the dataset that we are working with, we decided to perform several data preprocessing tasks to 1) narrow down the features that we will be working with, 2) handle missing values that may hinder our modeling task, and 3) condense our dataset to improve our model's performance and efficiency.

Data Preprocessing

As previously mentioned, we performed various data preprocessing tasks to simplify and clean the large amount of data that we are working with. Our initial cleaning was narrowing down the 85 features that were present in our dataset. We took a number of steps to achieve this. First, we narrowed down the data to only include traffic stops taking place in the city of San Francisco, because even though it was the most frequent city to appear in this dataset, dropping rows with other cities considerably condensed our dataset. This also accomplished the goal of focusing our project on an area that was relevant to us as Berkeley students (San Francisco).

Subsequently, we dropped all columns that either only have null values or have more than 50 percent of the data missing. Furthermore, we identified columns that we deemed as irrelevant to our task at hand, especially those that have more to do with personal identifiers, as opposed to features that we are interested in exploring. Some of these columns include doj_record_id, person_number, unique_identifier, data_as_of, location (Note: this column is different from 'city'). We also dropped the column perceived_age because not only are we more interested in the column perceived_age_group, we figured it would be easier to create categorical dummy variables for the latter.

Upon further analysis of the number of unique values of each column, we found more issues to address about the dataset. Firstly, duration_of_stop has too many unique values for one-hot encoding to be viable, so we separated those values into three buckets: 0 to 30 minutes, 30 to 60 minutes, and more than 60 minutes. We also decided it was necessary to one-hot encode the *district* column and the *traffic_violation_type_code* column. We also dropped columns containing information that would be collected after a search is already conducted because they are not relevant to our task.

Most importantly, we identified the relevant column that would act as our dependent variable for our predictive task, namely the 'actions_taken_code' column. It is important to note that there is an actions_taken column which describes by textual description what actions resulted from a particular traffic stop, and a corresponding actions_taken_code that listed the respective codes of those actions. The table in the Appendix describes the possible actions and their corresponding codes. With that established, we decided to only work with the numerical codes in the actions_taken_code column as opposed to the actions_taken column itself for ease of using RegEx filtering.

Specifically, we created a new column called "Search conducted," serving as a binary variable, indicating whether or not a person or property was searched as a result of the traffic stop. A result of 0 indicates that the search was not conducted, while a result of 1 indicates that the search was conducted. We achieved this task by checking, through RegEx, whether or not each item in the actions_taken_code column contained ‘18’, which corresponds to “Search of person was conducted,” or ‘20’, which corresponds to “Search of property was conducted.” However, before performing this filtering task, we made sure to drop rows that contained null values in the actions_taken_code column, as they essentially contained no useful information. This new “Search conducted” column will serve as our dependent variable of which we will perform our predictive task upon.

Our Model

We started out by running **OLS** and found a relatively low R^2 value of .262. From here, we calculated VIF to identify problematic features and found that there are some features with very high VIF values— for example the duration of the stop features each had “inf” as their VIF. However, we identified via the correlation matrix that when duration of stop was 30-60 seconds or greater than 60 seconds, it was highly correlated with a stop being conducted.

From here, we decided to move directly to **logistic regression** to first test out whether our OSR^2 would be higher for a model without much feature engineering. We got an OSR^2 of .2389 for the first logistic regression model we used with 40 features. Based on high p-values, we decided to drop 'longitude', 'supervisor_district', 'month', and 'if_k12_school_is_student_True'. We also dropped the duration of stop columns and Southern, Central, and Mission districts based on high VIFs. Just as we identified in the OLS model, the duration of the stop had a high VIF but was certainly highly correlated with a search being conducted because when we did attempt to drop these features, the OSR^2 decreased significantly to .1728. After dropping the features with high p-values, the OSR^2 did not increase by a significant amount. We also used the confusion matrix and found that our updated logistic regression model had an accuracy of .8056, as compared to the previously calculated baseline accuracy of .7897. The second logistic regression model with dropped duration and district features had a very low TPR of 20%, indicating that those features were important to include in future models.

After trying OLS and logistic regression, we decided to try a **random forest classifier** model to see if we could get a more accurate model. Before implementing the model, however, we discovered that we had two redundant ‘year’ columns and dropped one of them. We also decided to drop all the districts this time due to relatively high VIFs for all of them. From here, we ran the random forest classifier without validation and found a training accuracy of .8380 and a testing accuracy of .8290. We also had a TPR of .449, which was higher than that of the logistic regression model, but our FPR was also .004 higher. From here, we conducted 5-fold cross-validation with 500 n_estimators and were able to get a cross-validated accuracy of 0.8251.

We ran this validated classifier on the test set and got a test accuracy of 0.8255. It's also worth mentioning that because of how much data we had, we decided to validate on a smaller sample of our data (10%) to decrease runtime. It's possible that if we had the capability to validate on more of the data, the validated accuracy could be marginally different.

Impact and Conclusion

In recent years, police departments across the nation have increased their usage of predictive analytics in an effort to improve allocation of police resources and more effectively combat crime. If officers can accurately predict which individuals are highly likely to need a search, they can better manage their time and prioritize searches when deemed most necessary, resulting in a smoother experience for both officers and civilians being searched. In theory, our project would be able to identify the factors most predictive of a search being conducted and help police departments. However, it is extremely important to understand that in reality, just because certain factors have been predictive of a stop and search in the past, does not necessarily mean that those should be the same factors moving forward, especially if those factors are driven by biases within the system. In other words, a model like the one we developed can perpetuate biases hidden in its training data. When conducting EDA on our model, we noticed that demographic factors such as being African American or being a male were highly correlated with a search being conducted. Observations like these should be examined closely when evaluating whether or not a predictive policing model should be put into use. Identifying and addressing these biases can lead to fairer policing and searching practices by law enforcement.

Looking ahead, we see several opportunities to extend this analysis. We could enhance our model through advanced feature engineering, potentially uncovering more complex patterns that could improve predictive performance. Exploring other models, such as Gradient Boosting Machines or deep learning approaches, and implementing other ensemble models could also result in better accuracy and robustness. Expanding the parameter grid in future GridSearchCV iterations could also help in discovering optimal model settings, though this would increase computational demands. Exploring these options could significantly enhance the model's practical applicability and effectiveness in traffic stop incident analysis.

Appendix

Code	Action Taken
1	Person removed from vehicle by order
2	Person removed from vehicle by physical contact

3	Field sobriety test
4	Curbside detention
5	Handcuffed or flex cuffed
6	Patrol car detention
7	Canine removed search
8	Firearm pointed at person
9	Firearm discharged or used
10	Electronic device used
11	Impact projectile discharged or used
12	Canine bit or held person
13	Baton or other impact weapon used
14	Chemical spray use
15	Other physical or vehicle contact
16	Person photographed
17	Asked for consent to search person
18	Search of person was conducted
19	Asked for consent to search property
20	Search of property was conducted
21	Property was seized
22	Vehicle impound
23	Admission or written statement obtained from student
24	None

2. Features used for OLS (36 features): "duration_of_stop: 0-30", "duration_of_stop: 30-60", "duration_of_stop: 60>", "traffic_violation_type_code", "reason_for_stop_code", "year", "month", "day", "hour", "is_stop_response_to_call", "had_limited_or_no_english", "district_BAYVIEW", "district_CENTRAL", "district_INGLESIDE", "district_MISSION", "district_NORTHERN", "district_PARK", "district_RICHMOND", "district_SOUTHERN", "district_TARAVAL", "district_TENDERLOIN" , "perceived_race_ethnicity_Black/African American", "perceived_race_ethnicity_Hispanic/Latino(a)", "perceived_race_ethnicity_Middle Eastern or South Asian", "perceived_race_ethnicity_Multi-racial", "perceived_race_ethnicity_Native American", "perceived_race_ethnicity_Pacific Islander", "perceived_race_ethnicity_White", "perceived_gender_Gender Nonconforming", "perceived_gender_Male", "perceived_gender_Transgender man/boy", "perceived_gender_Transgender woman/girl", "perceived_age_group_30 - 39", "perceived_age_group_40 - 49", "perceived_age_group_50 - 59", "perceived_age_group_60 or over", "perceived_age_group_Under 18", "if_k12_school_is_student_True"

3. Features used for Logistic Regression Model 1 (36 features): "duration_of_stop: 0-30", "duration_of_stop: 30-60", "duration_of_stop: 60>", "traffic_violation_type_code", "reason_for_stop_code", "year", "month", "day", "hour", "is_stop_response_to_call", "had_limited_or_no_english", "district_BAYVIEW", "district_CENTRAL", "district_INGLESIDE", "district_MISSION", "district_NORTHERN", "district_PARK", "district_RICHMOND", "district_SOUTHERN", "district_TARAVAL", "district_TENDERLOIN", "perceived_race_ethnicity_Black/African American", "perceived_race_ethnicity_Hispanic/Latino(a)", "perceived_race_ethnicity_Middle Eastern or South Asian", "perceived_race_ethnicity_Multi-racial", "perceived_race_ethnicity_Native American", "perceived_race_ethnicity_Pacific Islander", "perceived_race_ethnicity_White", "perceived_gender_Gender Nonconforming", "perceived_gender_Male", "perceived_gender_Transgender man/boy", "perceived_gender_Transgender woman/girl", "perceived_age_group_30 - 39", "perceived_age_group_40 - 49", "perceived_age_group_50 - 59", "perceived_age_group_60 or over", "perceived_age_group_Under 18", "if_k12_school_is_student_True"

4. Features used for Logistic Regression Model 2 (28 features): "traffic_violation_type_code", "reason_for_stop_code", "day", "hour", "had_limited_or_no_english", "district_BAYVIEW", "district_INGLESIDE", "district_NORTHERN", "district_PARK", "district_RICHMOND", "district_TARAVAL", "district_TENDERLOIN", "perceived_race_ethnicity_Black/AfricanAmerican", "perceived_race_ethnicity_Hispanic/Latino(a)", "perceived_race_ethnicity_MiddleEasternorSouthAsian", "perceived_race_ethnicity_Multi-racial", "perceived_race_ethnicity_NativeAmerican", "perceived_race_ethnicity_PacificIslander", "perceived_race_ethnicity_White", "perceived_gender_GenderNonconforming", "perceived_gender_Male", "perceived_gender_Transgenderman/boy", "perceived_gender_Transgenderwoman/girl", "perceived_age_group_30-39", "perceived_age_group_40-49", "perceived_age_group_50-59", "perceived_age_group_60orover", "perceived_age_group_Under18"

5. We one-hot encoded the following features: 'city', 'results_of_stop_code', 'stop_data_record_status_code', 'contraband_or_evidence_code', 'actions_taken_code', 'is_stop_response_to_call', 'district', 'perceived_race_ethnicity', 'perceived_gender', 'perceived_age_group', 'had_limited_or_no_english', 'if_k12_school_is_student', 'traffic_violation_type'