Modeling Police Stop Duration using the San Diego RIPA Dataset

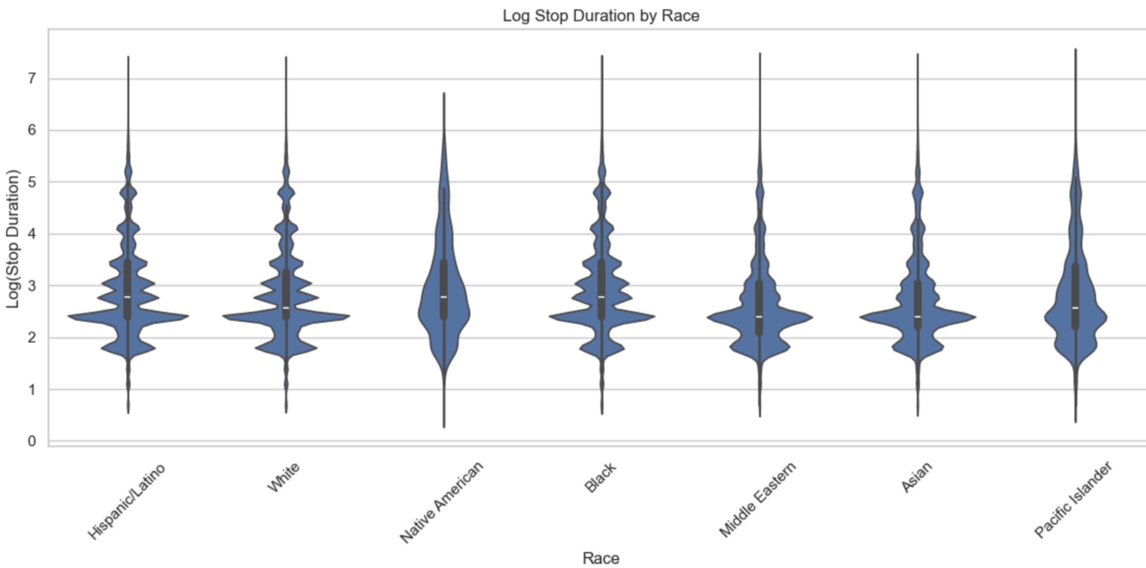
Harry Pachchala

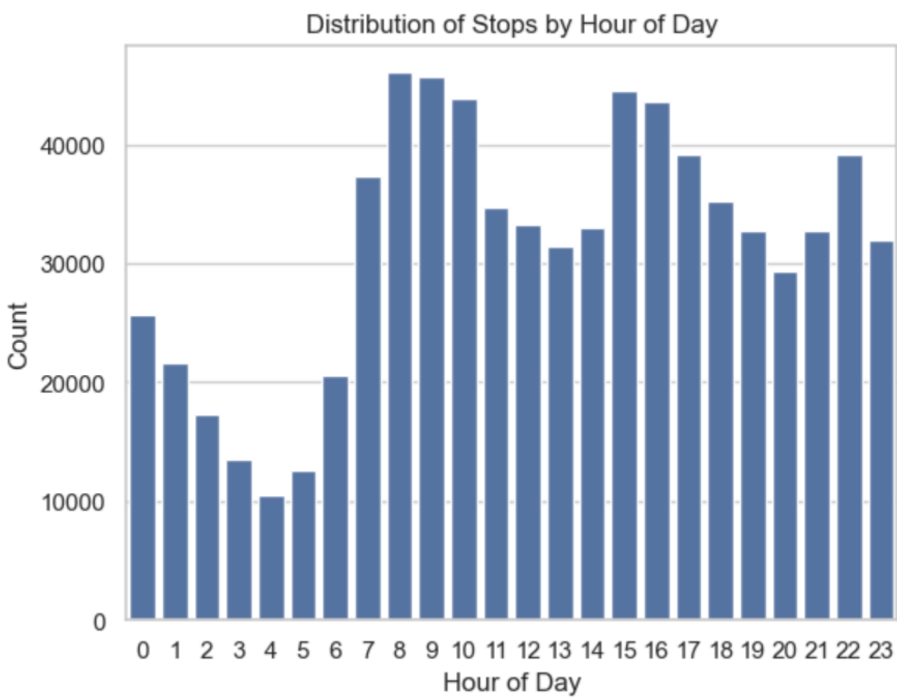
This project explores the factors that influence the length of police stops in San Diego using publicly available data from the Racial and Identity Profiling Act (RIPA). The goal was to build a regression model that predicts stop duration based on features such as stop reason, perceived demographics, and search context. More importantly, the project aimed to interpret what features most influence stop length and evaluate how well modern regression models can capture such patterns.

The data used in this analysis comes from the San Diego RIPA dataset. The base file, `ripa_stops_historic.csv`, was joined with 14 supplemental CSVs such as `ripa_gender_historic.csv`, `ripa_race_historic.csv`, `ripa_contraband_evid_historic.csv`, and others. These files include perceived race, gender, search justification, actions taken, and stop results. All were joined on the unique identifier `uid`, forming a wide and enriched dataset saved as `merged_ripa_dataset.csv`.

Once merged, the data was cleaned and prepared for modeling. Missing values in categorical fields were filled with the string `"Unknown"`, and numeric columns were imputed using the mean. `stop_time` was parsed into an hour field (`stop_hour`), and binned into time-of-day categories such as Morning, Afternoon, and Late Night. Categorical variables were one-hot encoded, and all numeric features were standardized using `StandardScaler`. Before modeling, a Random Forest model was used to rank features by importance, and only the top 50% of features were retained using `SelectFromModel`.

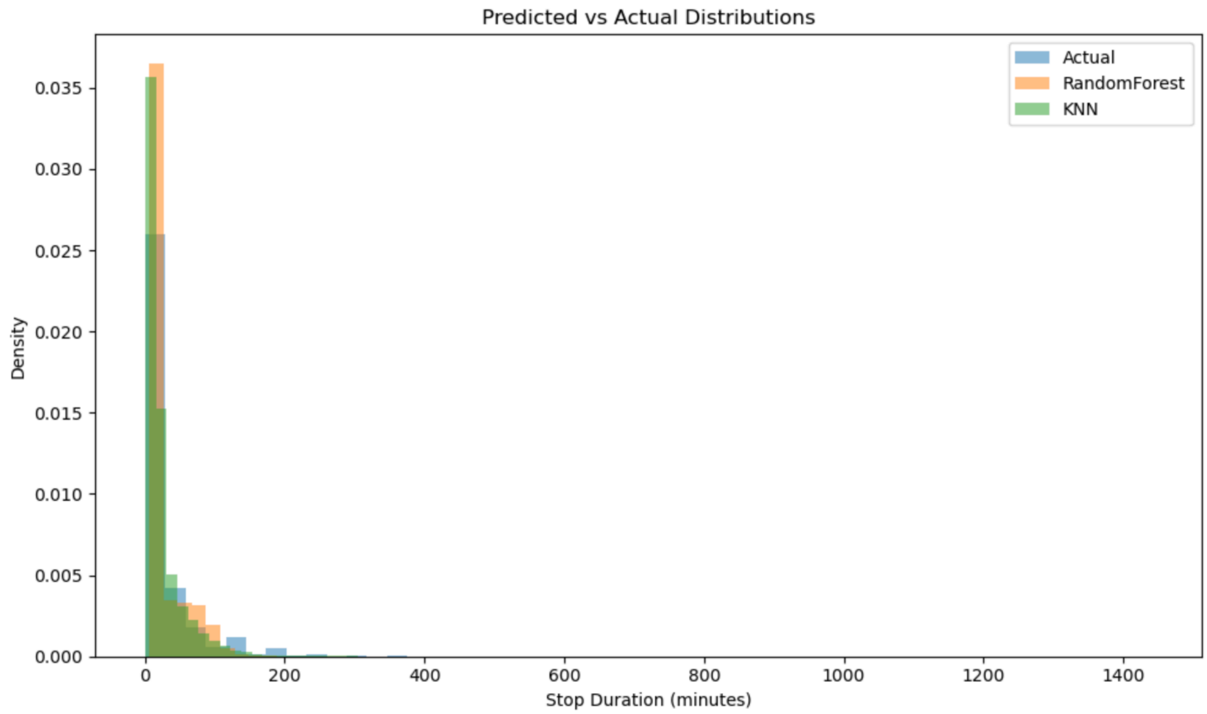
Exploratory analysis revealed that stop durations were heavily right-skewed, with the majority of stops lasting under 20 minutes. Stops involving searches, arrests, or the discovery of contraband tended to last significantly longer. There were visible patterns in stop length based on officer-perceived race and gender. For instance, perceived Hispanic/Latino individuals appeared overrepresented in longer stops, though causation cannot be inferred from these patterns alone. Time of day also showed subtle trends, with longer stops more common in the late evening and early morning hours.

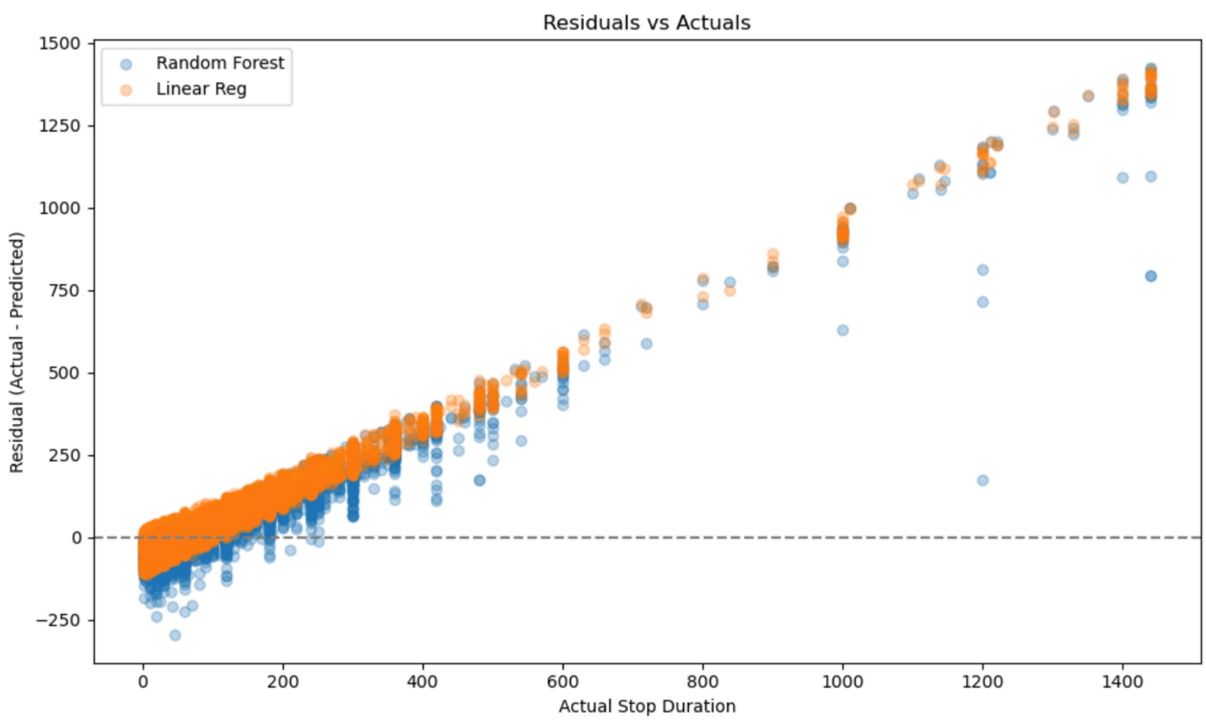




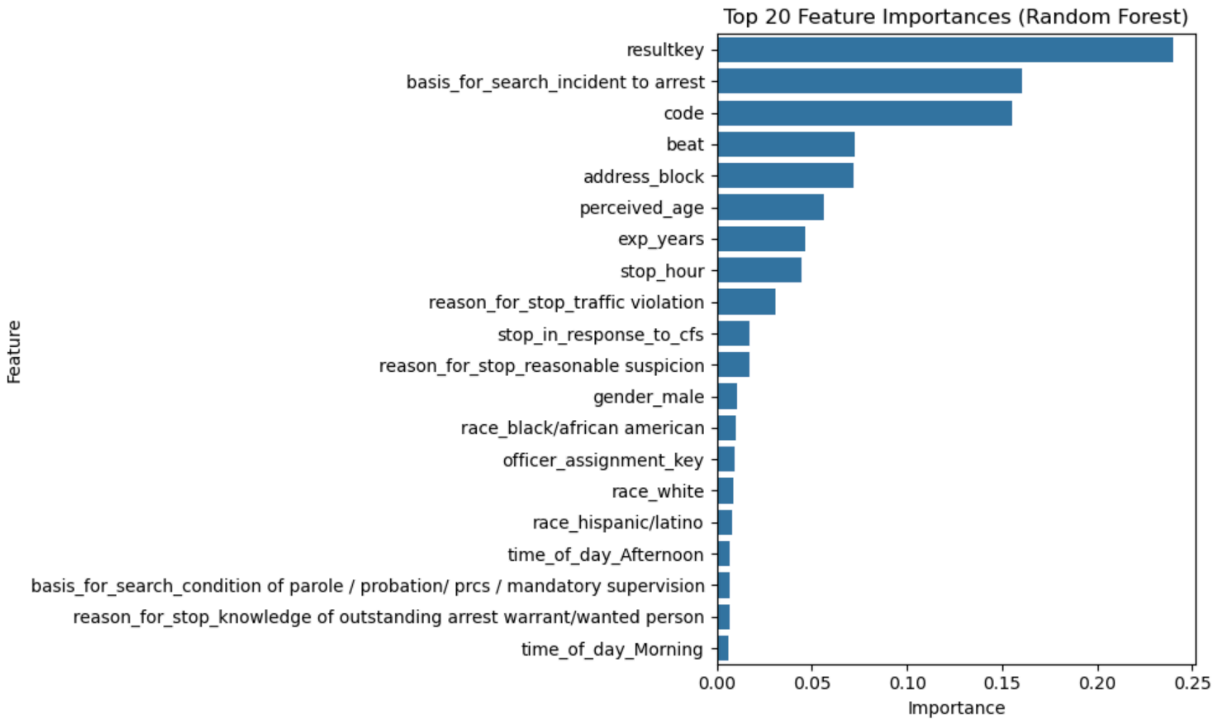
Several models were trained to predict stop duration, including a Dummy Regressor (as a baseline), Linear Regression, Ridge Regression with tuned alpha, a Random Forest Regressor, and a K-Nearest Neighbors (KNN) model trained on a 10,000-point sample. All models used the same set of pruned and scaled features. On the test set, Random Forest clearly outperformed the others, achieving an RMSE of 41.57 and an R^2 of 0.334. In contrast, both Linear and Ridge Regression achieved RMSEs of 45.56 and R^2 scores of 0.200. The KNN model performed worse, with an RMSE of 49.69 and an R^2 of just 0.049. The Dummy Regressor baseline had an RMSE of 50.95 and essentially zero R^2, confirming that all trained models added predictive value.

On the training set, Random Forest reached an RMSE of 35.65 and an R^2 of 0.489. While this gap suggests modest overfitting, it remains within acceptable limits for this kind of real-world data. To confirm stability, the Random Forest model was bootstrapped over 20 samples. The resulting RMSE scores ranged narrowly from 41.67 to 42.02, and R^2 ranged from 0.320 to 0.331. This reinforces the conclusion that the model is both relatively stable and generalizable.





The Random Forest's top features included `basis_for_search_incident`, `exp_years`, and perceived demographics such as `perceived_age`. These variables align with intuition and suggest that both procedural context and perceived identity play a role in stop length.



There are several limitations to the modeling. Most notably, `stopduration` is noisy and may be influenced by factors not present in the dataset, such as officer discretion, traffic conditions, or environmental context. Perceived demographics are recorded based on officer judgment, not self-report, which could introduce bias or inconsistency. Finally, model accuracy is inherently limited when trying to predict a highly variable, human-driven process.

Despite these challenges, the modeling pipeline demonstrates clear value. The results suggest that it is possible to identify meaningful patterns in stop duration, even if precise prediction remains difficult. A practical application of the model could be to flag unusually long stops given certain low-risk features, prompting review by oversight bodies. The feature importance rankings also offer transparency into which variables matter most.

For future work, this project could be extended by reframing the problem as a classification task, for example, predicting whether a stop will exceed a certain threshold (e.g., 30 minutes). Geographic clustering and visualization could also offer insight into location-specific practices. Finally, fairness audits (e.g., checking model performance across racial or gender groups) would be a valuable addition to assess equity in stop outcomes.