The Impact of New Bike Lanes on Urban Transportation Dynamics

Shaya Engelman shaya.engelman82@spsmail.cuny.edu

***Abstract*—**As urban populations grow, cities face challenges like congestion, sustainability, and safety. This study examines the impact of new bike lanes on bike usage, car dependency, and pedestrian safety, using data from bike-sharing programs in New York City and San Francisco, alongside traffic and accident data. The analysis employs statistical models such as logistic regression, Decision Trees, Random Forests, and XGBoost to assess the effectiveness of bike lanes. Findings show that new bike lanes increase bike usage, reduce car dependence, and improve pedestrian safety. The study highlights the importance of context-specific interventions, like protected lanes and traffic calming measures. The results provide actionable insights for city planners, recommending targeted bike lane expansions in low-traffic areas and complementary safety measures in high-traffic zones. This research offers valuable guidance for optimizing transportation networks and promoting sustainable urban mobility through enhanced cycling infrastructure.

1 INTRODUCTION

1.1 Background

As urban populations continue to surge, cities around the world are grappling with multifaceted challenges related to traffic congestion, economic productivity, environmental sustainability, and public safety. With more individuals residing in urban areas than ever before, the traditional reliance on motor vehicles has become increasingly untenable and changes are necessary. Many cities opt to attempt to transition more residents towards using bicycles as a mode of transportation. This

study seeks to address these challenges by examining the impact of new bike lanes on urban transportation dynamics, specifically focusing on their effects on bike usage, car usage, and pedestrian safety.

The hypothesis posits that adding new bike lanes will lead to increased bike usage, decreased car dependency, and reduced pedestrian accidents, thus contributing to less congestion and enhanced regional productivity. Given the multitude of factors influencing productivity, directly measuring the impact of bike lane changes on it is extremely challenging. Instead, this study accepts the findings by Hartgen and Fields (2009), which concluded that mitigating congestion can significantly bolster economic performance, as a given. Consequently, the focus is on finding increased bike usage as a key indicator. By examining these factors, this study aims to highlight the broader impacts of new bike lanes on urban environments.

The importance of bike usage cannot be overstated. Biking is not only an environmentally friendly mode of transport, but it also promotes healthier lifestyles among urban residents. As cities face rising pollution levels and public health crises related to sedentary lifestyles, increasing bike usage presents a compelling solution. Furthermore, the proliferation of bike-sharing networks, such as Citi Bike and Bay Wheels, has made biking more accessible and convenient than ever. These systems lower the barriers to entry for potential cyclists, offering a flexible and cost-effective alternative to car travel.

However, the financial implications of constructing bike lanes are considerable. Cities must navigate the complexities of funding such projects, often facing competition for limited municipal budgets. The costs associated with building and maintaining bike lane infrastructure must be weighed against the potential long-term benefits, including reduced traffic congestion and improved public safety. Additionally, the spatial allocation of roadways poses a challenge; dedicating space to bike lanes can result in the reduction of car lanes or parking spots. Urban planners must carefully balance these competing needs, ensuring that the overall transportation network remains efficient and effective.

1.2 Significance

The significance of this research lies in its potential to inform urban planning and policy decisions that can lead to substantial improvements in urban living conditions. As cities evolve, addressing traffic congestion is paramount, as it not only hampers economic productivity but also negatively impacts public safety and environmental quality. Hartgen and Fields (2009) demonstrated that enhancing travel speeds can lead to regional productivity gains of up to 1%, translating into meaningful economic benefits for urban areas. The link between shorter commutes and economic vitality is further reinforced by research from Prud'homme and Lee (1999), which highlights that cities with reduced commute times experience heightened productivity.

Moreover, the need for fewer pedestrian and cyclist accidents is critical in today's urban landscapes. As bike usage rises, so too does the potential for conflicts between cyclists, pedestrians, and motor vehicles. Well-designed bike lanes can help mitigate these risks by creating dedicated spaces for cyclists, thereby reducing accidents, and promoting safer road environments. This aspect of urban safety is not merely a matter of public health; it also affects the economic landscape. Increased safety leads to greater confidence in cycling as a viable transportation option, further encouraging its adoption.

In light of the COVID-19 pandemic, which prompted rapid, diverse responses in over five hundred cities, the need for innovative transport solutions has never been more pressing. Many urban areas have reallocated street spaces to support walking, cycling, and outdoor commerce. The emergence of these measures underscores the importance of dedicated cycling infrastructure in urban planning, particularly in fostering resilient, adaptive transportation networks that can respond to changing mobility demands (Combs & Pardo, 2021).

Ultimately, this study aims to provide critical insights for city planners and policymakers. By evaluating the trade-offs involved in constructing bike lanes—including initial financial costs, space allocation, and the potential for reduced car dependency—this research will illuminate the long-term benefits of improved safety, reduced commute times, and enhanced economic productivity. As cities continue to navigate the complexities of urbanization, understanding the role of

bike lanes in transportation dynamics will be essential in fostering sustainable urban growth.

Understanding the causal relationships between bike lane additions and accident rates is crucial for informing evidence-based policy decisions. By investigating these dynamics, this research aims to provide insights that not only contribute to academic discourse but also offer practical guidance for urban planners and policymakers seeking to enhance cycling safety and infrastructure in their cities.

2 PRIOR RESEARCH

The literature on the impact of bike lanes on urban transportation presents a varied landscape, with some studies indicating a positive relationship between bike lane infrastructure and cycling rates, while others report minimal effects. For instance, Hwang et al. (2023) found a significant increase in non-motorized transportation methods in urban areas but did not observe a corresponding rise in biking specifically. Similarly, Buck et al. (2011) identified a strong correlation between the presence of bike lanes and the proximity of bike share stations, suggesting that infrastructure improvements can drive bike-sharing utilization.

Research conducted by Hartgen and Fields (2009) indicated that alleviating traffic congestion through improved travel speeds could enhance regional productivity by up to 1%. This is further supported by Prud'homme and Lee (1999), who demonstrated that shorter commute times contribute significantly to economic efficiency. As such, this study aims to explore whether the installation of bike lanes can enable cities to achieve these economic benefits by alleviating gridlock and optimizing commute times.

One of the most significant contributions to this field comes from Kraus and Koch (2021), who studied the impact of provisional bike lane infrastructure introduced during the COVID-19 pandemic. Their research revealed that temporary bike lanes led to substantial increases in cycling, with rates rising between 11% and 48% in cities that implemented these lanes. This demonstrates the potential for even short-term infrastructure changes to encourage long-lasting shifts in transportation behavior.

Additionally, Arancibia et al. (2019) studied the economic impact of bike lanes in Toronto and found that replacing on-street parking with bike lanes did not harm

local businesses. In fact, businesses saw increased customer spending and foot traffic after the lanes were installed. This finding challenges the perception that bike lanes detract from commercial activity, suggesting instead that they may stimulate economic growth by attracting more foot and bike traffic.

Despite the growing body of literature exploring the relationship between bike lanes and cycling rates, there remains a notable gap in comprehensive research regarding their effects on accidents in general, and, in particular, pedestrian accidents. Buehler and Dill (2015) highlighted that the introduction of bike lanes often leads to positive changes in urban dynamics, indicating that dedicated cycling space can reduce pedestrian injuries by improving overall traffic organization. This study seeks to fill this research gap, providing a holistic view of how bike lanes impact urban transportation dynamics, encompassing bike and car usage, pedestrian safety, and overall traffic patterns.

Expanding this research to multiple cities will yield more reliable and generalizable results. This multi-city approach will aim to identify consistent patterns across different urban environments, enhancing our understanding of bike lanes' impact. It is essential to address the reduction or smaller increase in car ridership. While projecting bike ridership growth is useful, comprehending the shift from car users to bike users is equally critical. This study will specifically investigate this aspect, offering valuable insights into how bike lanes can effectively reduce car usage and promote sustainable urban transportation.

3 METHODOLOGY

3.1 Data Collection

3.1.1 Bike Usage Data

The study utilized bike-sharing data from three cities: Citi Bike in New York City, Bay Wheels in San Francisco, and Divvy in Chicago, all operated by Lyft. Monthly reports on bike-sharing usage were obtained for each city, with each month represented by a separate file. In the case of Chicago, some months were combined in the datasets, necessitating the splitting of these files to ensure monthly granularity. A dedicated dataframe was created for each city to compile the total bike-sharing counts per month.

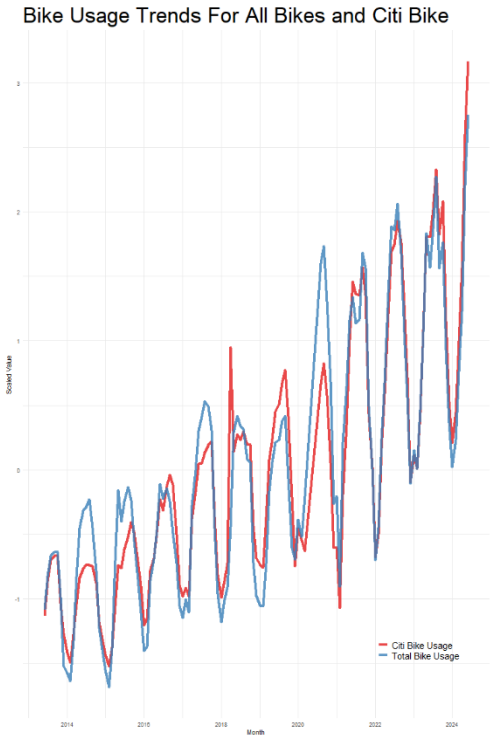


Figure 1; Citi Bike usage data started off with smaller fluctuations than total bicycle usage but as it gains traction it converges to match the range.

To enhance the accuracy of the analysis and address concerns regarding the representativeness of bike-sharing data, validation was performed using additional data sources. Automated bike count data, sourced from NYC's open data platform, was compared against the bike-sharing figures. The trends observed in both datasets converged to closely mirror each other, confirming that bike-sharing data serves as a valid proxy for overall bike usage in an urban environment. Consequently, this validation process was not repeated for SF and Chicago. Figure 1.

3.1.2 Traffic Count and Bike Lane Data

Traffic count data was obtained from the open data platforms of NYC and SF. Accurate traffic count data for Chicago could not be sourced. To quantify the extent of bike lane infrastructure, bike lane data from NYC was downloaded from which total bike lane footage was calculated using the installation dates. This data was grouped by month and integrated into the NYC dataframe. A similar process was employed for SF, where bike lane data was also aggregated by month. For Chicago, only partial bike lane data was available from two separate websites. This data, provided on a yearly basis, was manually processed, and added to the Chicago dataframe.

3.1.3 Demographic and Accident Data

Population estimates for each city were derived from U.S. Census data and incorporated into all three datasets. Additionally, pedestrian and cyclist accident data were collected from the respective open data platforms for each city. This

included total counts of pedestrian accidents, fatalities, cyclist accidents, and fatalities, resulting in six accident-related columns being added to each city's dataframe.

3.2 Data Preparation and Transformation

In preparation for analysis, the dataset underwent several transformations. For instance, the bike-sharing data was aggregated by month to allow for time-series analysis. In Chicago, some filenames in the dataset were renamed for consistency, utilizing a function that converted filenames into a standardized date format. This process ensured that data from all three cities could be accurately compared.

Missing values were identified and addressed through several approaches. An initial assessment quantified the extent of missing data across various variables, which was essential for understanding the dataset's completeness and integrity. Recognizing the potential impact of missing values on the analysis, particular attention was given to key metrics.

Data prior to the introduction of the Citi Bike network in June 2013 was removed. This step was crucial, as the establishment of the bike-sharing program significantly altered cycling dynamics in NYC. By filtering the dataset to include only relevant data, the analysis could more accurately reflect the effects of bike lanes and the bike-sharing system.

To handle missing values for specific variables, a multiple imputation technique was employed. This method helped fill gaps in the data for the total bike lane length variable, enhancing the robustness of the dataset. Additionally, linear regression models were used to impute missing values for total bike counts and traffic volume, ensuring that the imputed values were informed by existing data trends.

These comprehensive data cleaning and preparation steps ensured that the datasets were both complete and ready for rigorous analysis.

3.3 Data Analysis

Next, as a preliminary step to account for the relative size of cities on their transportation patterns, I analyzed bike usage as a proportion of motor vehicle

traffic. This metric aimed to provide a standardized measure of bike usage, allowing for more meaningful comparisons across cities. The cumulative sum of bike lane additions was also calculated to assess the impact of infrastructure changes on bike-to-traffic ratios over time.

Since bike usage patterns can be severely impacted by seasonal variations, the time series data was decomposed to isolate the trend and seasonality components. This decomposition process helped identify underlying patterns in bike usage and traffic volume, enabling a more nuanced analysis of the data. After removing the seasonal component, the trend of bike-to-traffic ratios was plotted against the trend of cumulative bike lane additions to explore potential relationships between these variables.

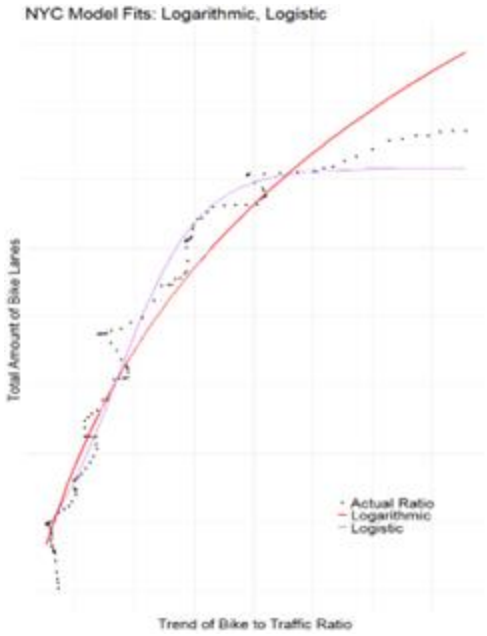


Figure 2; Logistic Regression seems to fit the actual NYC data best, but Logarithmic Regression is likely better for extrapolating future points

In the NYC dataset, this visual analysis revealed a strong resemblance to a logistic curve, suggesting a saturation point for bike lane additions. However, the data also exhibited a slight upward trend after the inflection point, instead of the typical flattening seen in logistic curves. Additionally, a logistic curve is an illogical fit for bike lane additions, as it is unlikely for these to be a strict saturation point where no more bike lanes can be added, while the effect slowing down as more bike lanes are added is more likely. After further exploration, a logarithmic model was considered as a potentially better fit for the data as it too fit the data reasonably well while also allowing for the continuing upward trend in the data. Figure 2.

Building on these visual observations, a series of regression models were developed to explore potential relationships between bike lane additions, bike-to-

traffic ratios, and safety outcomes. The preliminary results of these exploratory models are summarized in Table 1, highlighting potential trends rather than definitive conclusions. The Log Lane-Bike Ratio Model, which plots cumulative bike lane additions against the logarithm of the bike-to-traffic ratio trend, showed a very strong fit (R-squared = 0.9426). This suggests that 94.26% of the variance in cumulative bike lane additions can be explained by the model, with highly significant coefficients (p-value < 2e-16). This statistical significance confirms a strong relationship between bike lane additions and the log-transformed bike-to-traffic ratio, reinforcing the infrastructure's impact on bike usage trends.

In contrast, the Linear Bike Ratio-Injuries Model, which analyzed the relationship between the bike-to-traffic ratio and total injuries and deaths, had a low R-squared value (0.01389), meaning the bike-to-traffic ratio explained little of the variance in injuries and deaths. The non-significant p-value (0.183) further suggests that the relationship is not statistically meaningful. Given the small effect size of the bike-to-traffic ratio relative to overall traffic, it's likely that the ratio alone is insufficient to explain variations in injury rates.

The Linear Traffic-Injuries Model demonstrated a more substantial relationship between motor vehicle traffic volume and injuries and deaths, with an R-squared of 0.1403. Although this explains only 14.03% of the variance, the highly significant p-value (1.22e-05) indicates a strong correlation between higher traffic volumes and increased accident rates.

While a direct statistically significant relationship between the bike-to-traffic ratio and total injuries and deaths was not identified, a clear positive relationship exists between motor vehicle traffic volume and total injuries and deaths. Consequently, there must be a connection between fewer cars on the road (indicated by a higher bike-to-traffic ratio) and fewer injuries and deaths. This relationship was not captured in the linear model likely due to the small effect size relative to total traffic volume. Establishing this connection, along with a positive logarithmic relationship between bike lane additions and the bike-to-traffic ratio (p-value < 2e-16, R-squared = 0.9426), infers that adding bike lanes correlates with fewer injuries and deaths.

Table 1; Summary outputs from the various regression models created. Note the high and statistically insignificant value in the second model.

Model	Term	Estimate	Standard Error	t-Value	P-Value	R-Squared
Log Lane-Bike Ratio Model	Log-Transformed Bike-to-Traffic Ratio	2.069258e+06	4.529407e+04	45.684966	1.07923649553373e-80	0.9426408
Bike Ratio-Injuries Model	Bike-to-Traffic Ratio	-4.641118e+01	3.470183e+01	-1.337428	0.183473386112087	0.0138887
Traffic-Injuries Model	Motor Vehicle Traffic Volume	5.170000e-04	1.136000e-04	4.552947	1.22288388016561e-05	0.1403196

A similar approach was utilized for the San Francisco data. First a variety of patterns were visualized to determine if the data easily fits any of the common regression lines. As with the NYC data, both a logarithmic and logistic regression line were relative fits. However, neither of them fit as closely as the NYC data did. For the San Fracisco data, the logarithmic model seems to be a better fit than the logistic one, further reinforcing that the logarithmic model is a better starting point for analysis. Figure 3.

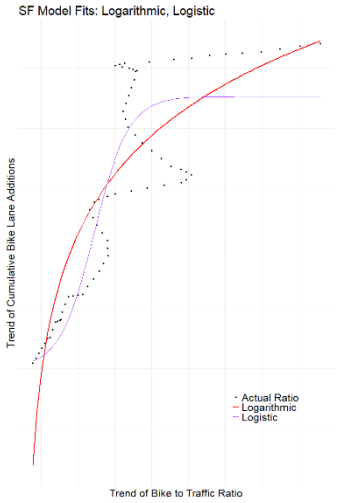


Figure 3

A similar approach was utilized for the San Francisco data. First, a variety of patterns were visualized to determine if the data easily fit any of the common regression lines. As with the NYC data, both a logarithmic and logistic regression line were reasonable fits. However, neither of them fit as closely as the NYC data did. For the San Francisco data, the logarithmic model emerged as the better fit compared to the logistic model, further reinforcing that the logarithmic model is a more suitable starting point for analysis.

The Log Lane-Bike Ratio Model, which analyzed cumulative bike lane additions against the logarithm of the bike-to-traffic ratio, yielded a strong fit (R-squared = 0.714). This indicates that 71.4% of the variance in cumulative bike lane additions can be explained by the logarithmic transformation of the bike-to-traffic ratio. The statistically significant coefficient (p-value < 2e-16) underscores the robustness of this relationship, highlighting the importance of infrastructure changes in shaping bike usage patterns.

In contrast, the Linear Bike Ratio-Injuries Model examined the relationship between the bike-to-traffic ratio and total injuries and deaths. This model revealed

a moderate fit, with an R-squared value of 0.1837, suggesting that the bike-to-traffic ratio explains only 18.37% of the variance in injuries and deaths. The significant p-value (0.000192) indicates a meaningful relationship, pointing to the potential influence of increased bike usage on safety outcomes.

Finally, the Linear Traffic-Injuries Model analyzed the impact of motor vehicle traffic volume on total injuries and deaths. Although the model showed a very low R-squared value (0.02044), suggesting limited explanatory power, it indicated that motor vehicle traffic has a strong correlation with injuries, as evidenced by the significant intercept (p-value < 2e-16) and a coefficient indicating a small effect size for traffic volume.

Overall, while these preliminary models highlight different aspects of the relationship between bike usage and safety, they collectively suggest that increasing bike infrastructure may contribute positively to reducing injuries and deaths in San Francisco; however, further analysis is necessary to draw definitive conclusions. The findings emphasize the complexity of urban cycling dynamics and the need for targeted interventions to enhance safety while promoting bike usage.

4 NEXT STEPS

The preliminary regression models provided valuable insights but were not conclusive, indicating that further, more robust analyses are necessary. They reveal a statistically significant relationship but not exactly what it is, it isn't linear or strictly logarithmic either. Additionally, while being statistically significant, the effect size of bike lane additions on accidents is relatively small, making it harder to model as well. However, these models served as a great exploratory analysis into what the next steps of the research should be.

4.1 Decision Trees: Uncovering Nonlinear Interactions and Thresholds

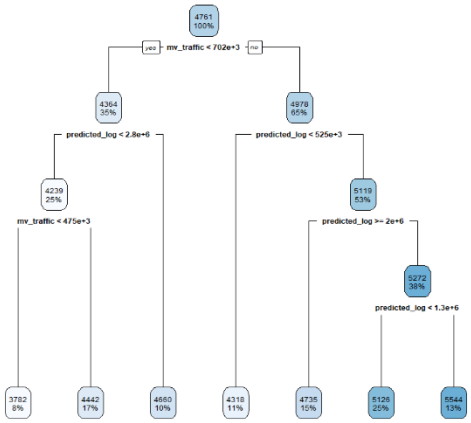
The decision tree model was a critical choice for exploring non-linear patterns and thresholds within the dataset, especially given the small size of fewer than 200 records. This limited dataset posed a significant risk of overfitting, where the model might capture noise rather than true underlying relationships.

Consequently, a major focus of the tuning process was ensuring that the decision tree achieved a balance between complexity and generalizability.

To address this, the decision tree parameters were carefully fine-tuned through a meticulous grid search. This process systematically tested combinations of parameters such as the complexity parameter (CP), maximum tree depth, and minimum split size to identify a configuration that minimized overfitting while retaining meaningful patterns. The final model, with a complexity parameter of 0.0301 and a maximum depth of 4, represented an optimal balance. These parameters ensured that the tree captured significant interactions, such as those between motor vehicle traffic volume (mv_traffic) and the logarithmic regression-derived variable (predicted_log) representing the expected trend of bike usage, without becoming overly tailored to the training data.

The predicted_log variable was derived from a logarithmic model, which was selected because the prior analysis indicated that the relationship between bike lane additions and the trend of the ratio of bike usage to motor vehicle usage. The predicted_log variable captures the cumulative bike lane development over time, accounting for the rate of expansion and its diminishing effect. By incorporating this variable into the Decision Tree, the model was able to assess how different stages of bike lane development—early expansion versus more established networks—interact with other factors, like traffic volume, to influence safety outcomes.

The model's performance metrics validated the chosen approach. The root node error was reduced from 1.00000 to 0.50416 after six splits, demonstrating the tree's ability to extract meaningful thresholds. Cross-validation metrics further confirmed the model's robustness, with an xerror of 0.84283 and a standard deviation of 0.100767, underscoring its stability and reliability against overfitting. The tuning process also considered the practical implications of the dataset's small size. By limiting the maximum depth to 4, the model avoided excessive branching, which could have resulted in highly specific splits that lacked predictive power for unseen data. This careful design ensured that the decision tree provided



interpretable results that policymakers could trust, identifying actionable thresholds where bike lane additions had the greatest impact.

The tree's splits revealed valuable insights for urban planners, identifying actionable thresholds where bike lane interventions could have the greatest impact. The root split occurred at an mv_traffic threshold of approximately 700,000 vehicles per month, separating high-traffic months, where bike lane additions had a limited safety effect, from lower-traffic months, where safety improvements were more pronounced. Subsequent splits incorporated the predicted_log variable, identifying zones of heightened injury risk with scores above 0.7. These areas demonstrated the most substantial benefits from targeted bike lane expansions.

Interactions uncovered by the model underscored the importance of context-specific interventions. For example, in high-traffic months exceeding the 700,000-vehicle threshold, the effect of bike lanes was diminished, suggesting that complementary measures like protected lanes or reduced traffic speeds are necessary. Conversely, low-traffic months with lower predicted_log scores showed significant reductions in injuries, affirming the efficacy of bike lane additions in these settings.

This detailed analysis provides clear, data-driven guidelines for policymakers:

- Prioritize bike lane additions in lower-traffic areas (<700,000 vehicles/month) where the impact on safety is maximized.
- Supplement bike lanes in high-traffic zones with additional safety measures to enhance their effectiveness.

Focus efforts on high-risk zones identified by the predicted_log variable, ensuring resources are directed where they can achieve the greatest safety improvements.

Figure 4

This balance between model complexity, interpretability, and generalizability ensures that the decision tree analysis not only reflects the dataset's constraints but also yields actionable insights for policymakers navigating urban safety challenges. The decision tree serves as a foundational tool for any future planning, identifying where bike lane additions are most likely to have a significant impact on safety, allowing urban planners to prioritize interventions that maximize their effectiveness. Other models will be explored, but the decision tree remains unparalleled in simple and interpretable actionable insights.

4.2 Random Forest

The random forest model was a logical next choice for this analysis, as it provides a more robust and reliable alternative to a single decision tree. A major limitation of the data was its small size, covering only around a decade of data, leaving any model susceptible to overfitting. The key advantage of random forest lies in its ensemble nature, where multiple decision trees are aggregated to make predictions, helping to reduce the risk of overfitting and capturing complex patterns within the data. While larger numbers of trees can often lead to marginally better performance, a smaller number of trees was carefully selected in this case. This decision was made to keep the model simpler and more interpretable, especially considering the limited size of the dataset, thereby avoiding the risk of overfitting to noise.

To ensure that the random forest model was both accurate and generalizable, the hyperparameters for this model too were finely-tuned through a meticulous grid search and careful pruning. The grid search process tested different combinations of parameters, including the number of trees, the number of variables tested at each split, and the maximum depth of the trees. After careful consideration of the trade-offs between performance and model complexity, a final configuration of 37 trees with one variable tested at each split was chosen. This model configuration

balanced accuracy and simplicity, ensuring that the model captured significant patterns without becoming overly complex and tailored to the training data.

The performance of the model was reflected in a mean squared residuals (MSE) value of 0.5676275, which indicates how well the model predicted the outcome variable. Additionally, the model explained 41.42% of the variance in the data, showing that it successfully captured a substantial portion of the underlying relationships in the dataset.

The variable importance analysis revealed the key predictors that were driving the model's predictions:

- predicted_log: With an importance score of 32.32, this variable, derived from the logarithmic regression model, proved to be a crucial factor in determining safety outcomes, highlighting the ongoing impact of cumulative bike lane development.
- mv_traffic: With an importance score of 31.87, motor vehicle traffic volume played a critical role in shaping the safety outcomes. High traffic volumes often reduce the effectiveness of bike lane interventions, making it an essential factor in the model.
- cumulative_bike_lane: With an importance score of 28.80, this variable, reflecting the extent of bike lane development, showed its significant role in understanding the relationship between bike lane presence and safety outcomes.

These results show that the random forest model successfully captured meaningful interactions in the data. The predicted_log and mv_traffic variables, in particular, were identified as key drivers of safety outcomes, underlining the importance of both bike lane development and traffic volume in shaping urban safety.

Despite the relatively simple model configuration, the random forest model proved to be highly effective at identifying actionable insights. For example, in areas with high predicted_logistic scores (indicating high rates of bike lane expansion) and moderate mv_traffic levels, additional bike lane development

could significantly improve safety outcomes. In contrast, high-traffic areas may require additional measures, such as protected bike lanes or reduced traffic speeds, to enhance safety.

In conclusion, the random forest model, with its carefully pruned structure and optimized hyperparameters, complements the decision tree analysis by offering a more robust and generalizable framework for urban planning. By balancing model complexity and accuracy, it provides actionable insights that can guide bike lane interventions and improve urban safety planning.

4.3 XGBoost: Leveraging Advanced Ensemble Learning for Improved Predictive Accuracy

The XGBoost model was another crucial tool in this analysis, chosen for its ability to combine the power of gradient boosting with regularization techniques to enhance predictive performance. XGBoost, a highly efficient implementation of gradient boosting, is known for its robustness and ability to handle complex relationships in data, making it particularly well-suited for datasets with intricate patterns and potential interactions. This model was chosen to further refine the predictions and assess whether an advanced ensemble technique could offer superior performance compared to the simpler Decision Tree and Random Forest models.

XGBoost works by constructing a series of decision trees in a sequential manner, where each tree attempts to correct the errors made by the previous one. This iterative process allows XGBoost to capture even the most subtle relationships between variables, leading to higher accuracy and better generalization. Regularization in XGBoost, which penalizes overly complex trees, helps mitigate the risk of overfitting—a key concern given the small size of the dataset.

For this analysis, the XGBoost model was carefully tuned using parameters that optimize both training speed and model performance. The training process utilized 100 iterations, with early stopping implemented after 10 rounds to prevent overfitting and ensure the model did not continue training once it reached optimal performance. The model was trained with an objective of reg:squarederror,

appropriate for regression tasks, and evaluated using the root mean squared error (RMSE) metric to assess predictive accuracy. As shown in the evaluation log, the RMSE steadily decreased, with the model achieving an RMSE of 0.0016395 after the final iteration, indicating exceptional precision in predicting the outcome variable.

One of the key advantages of XGBoost is its ability to produce feature importance scores, which help interpret the model and understand the relative contribution of each predictor. The variable importance analysis for the XGBoost model highlighted the following key features:

- predicted_log: With a Gain score of 0.4876, the predicted_log variable was by far the most influential predictor in the model, capturing the cumulative impact of bike lane development over time. This variable continued to show its relevance, consistent with its role in both the decision T=tree and random forest models.

- mv_traffic: This variable, representing motor vehicle traffic volume, came in second with a Gain score of 0.3390. As in the previous models, traffic volume played a crucial role in shaping the safety outcomes, emphasizing the importance of considering traffic conditions when planning bike lane interventions.

- cumulative_bike_lane: The third most important feature was the cumulative_bike_lane variable, with a Gain score of 0.1733. While still significant, its importance was somewhat lower compared to the other two variables, reflecting its role in explaining the relationship between bike lane infrastructure and safety outcomes.

These feature importance scores align with the findings from the decision tree and random forest models, confirming that bike lane development (as captured by predicted_log and cumulative_bike_lane) and traffic volume (mv_traffic) are central factors influencing safety outcomes. However, the XGBoost model provided a more nuanced view of these relationships, further refining the understanding of how these variables interact.

The XGBoost model's performance was evaluated using the RMSE metric, and its predictions revealed similar patterns to the decision tree and random forest models. Notably, in areas with high predicted_log scores and moderate traffic volumes, the addition of more bike lanes showed substantial improvements in safety outcomes. However, as in the previous models, XGBoost suggested that areas with very high traffic volumes might require complementary safety measures, such as protected bike lanes or traffic speed reductions, to effectively reduce injuries.

The combination of advanced boosting techniques and regularization made XGBoost particularly effective at identifying subtle interactions between variables. For example, while the decision tree captured the interaction between mv_traffic and predicted_log with distinct thresholds, XGBoost was able to model these relationships in a more continuous, smoother way. This increased the model's ability to generalize across different urban settings, particularly in cases where data points might be sparse or noisy.

In conclusion, the XGBoost model provided valuable insights that complement the findings from the decision tree and random forest analyses. By capturing more complex interactions and delivering highly accurate predictions, XGBoost reinforced the importance of targeted bike lane interventions in areas with low to moderate traffic volumes and highlighted the need for additional safety measures in high-traffic zones. This model, with its emphasis on accuracy and generalization, serves as an essential tool for urban planners, offering a refined approach to making data-driven decisions for safer, more effective bike lane development.

This enhanced predictive framework, combining the strengths of decision trees, random forests, and XGBoost, offers a comprehensive methodology for optimizing urban infrastructure interventions. It ensures that resources are directed toward the areas where they will have the most significant impact on improving safety for cyclists, providing policymakers with the tools needed to create safer and more sustainable cities.

5 SAN FRANCISCO ANALYSIS

The same analysis was done on the San Francisco dataset. While some details like the best model settings and specific split points were different, the overall importance of the variables was very similar. This supports the earlier findings, showing that even with different model setups, the main factors influencing the predictions stayed the same across datasets. This suggests that the relationships identified are strong and can apply to different areas, confirming that the results are valid in various locations. The consistent importance of certain variables also highlights that they play a key role in the phenomena being studied, regardless of local differences in the data or model settings. Overall, these results give us more confidence that the model can make reliable predictions in different situations.

5.1 Decision Trees: Uncovering Nonlinear Interactions and Thresholds

The San Francisco dataset was even smaller than the NYC one. This makes the risk of overfitting even more prone and is more of a reason for more complex models to fail in adapting to unseen data. This lead to the strength of the decision tree and to the pruning to be done with extreme caution.

As done previously with the NYC data, hyperparameters were carefully tuned using grid search. Key parameters such as the complexity parameter (CP), maximum tree depth, and minimum split size were systematically tested. The final model, with a complexity parameter of 0.09 struck a balance between capturing meaningful patterns and avoiding overfitting. These parameters enabled the tree to identify significant interactions, such as those between motor vehicle traffic volume (mv_traffic) and the `predicted_log` variable, without sacrificing generalizability.

Figure 5

Model performance metrics confirmed the chosen configuration's effectiveness. The root node error decreased significantly, indicating the tree's ability to extract critical thresholds. Cross-validation results demonstrated robustness, with an

xerror of 0.812 and a standard deviation of 0.092. The model's limited depth prevented overfitting, ensuring interpretability—an essential characteristic for policy application.

The decision tree splits for the San Francisco model provided useful insights for urban planners. The root split showed that when the predicted_log value is greater than or equal to 175.0507, the average safety score was lower (around 280), while areas with a predicted_log value below 175.0507 had a higher average safety score (around 351). This suggests that areas with lower predicted_log values could benefit more from bike lane interventions. The model's simplicity highlights the importance of focusing on areas with specific characteristics as defined by the predicted_log threshold for effective safety improvements.

These findings emphasize the importance of context-specific strategies. In areas where the predicted trend of the car-to-bike ratio (represented by predicted_log) is high (above 175.0507), bike lanes alone may not be sufficient to improve safety, suggesting that additional measures like protected lanes or traffic calming are needed. In contrast, areas with a lower predicted car-to-bike ratio saw significant safety improvements from bike lane expansions, indicating that targeted interventions can be particularly effective in these areas.

Key Recommendations:

- Prioritize bike lane additions in areas with traffic volumes below 70,000 vehicles/month.
- Implement supplemental measures in high ratio of bike to motor vehicle months and zones to enhance safety.

The decision tree's optimized structure ensures it captures actionable patterns, making it a valuable tool for identifying critical thresholds where bike lane interventions are most impactful.

5.2 Random Forest: Aggregating Decision Trees for Robust Insights

While the output of the decision tree was very straightforward, random forest and XGBoost were utilized in the San Francisco analysis due to their strong

performance in the New York City (NYC) dataset, where they provided robust, generalizable insights into the impact of bike lane development on urban safety. However, in San Francisco, where the dataset was smaller and more sparse, these models underperformed relative to expectations. They were still valuable, but failed to exceed the value of the decision tree models.

As done previously with the NYC data, the random forest model leveraged its ensemble nature to mitigate overfitting and capture more complex patterns in the data. By aggregating predictions from multiple trees, random forest aimed to provide reliable insights while maintaining generalizability. However, due to the limited dataset size in San Francisco, the model's ability to uncover deep interactions was restricted.

Hyperparameters were fine-tuned using grid search to balance model performance and complexity. Key parameters such as the number of trees, the number of variables considered at each split, and the maximum tree depth were tested systematically. The final configuration of 18 trees, with two variables considered per split, ensured the model could identify significant patterns without being overly complex or tailored to the training data.

The model's performance was moderate, with a mean squared error (MSE) of 4111.169 and an explained variance of just 1.87%, suggesting limited predictive power. Despite this, variable importance analysis revealed several key insights:

- `predicted_log`: With the highest importance score of 90283.79, this variable was the most influential, confirming its critical role in understanding the predicted car-to-bike ratio and its impact on bike lane safety outcomes.
- `mv_traffic`: Scoring 83258.37, this variable represents the extent of bike lane infrastructure and how it influences urban safety, though its impact was somewhat lower than that of predicted_log.
- `cumulative_bike_lane`: With an importance score of 55596.84, traffic volume played a consistent role in shaping safety outcomes, underscoring how bike lane effectiveness varies depending on traffic density.

While the Random Forest model showed moderate performance overall, it successfully highlighted important patterns, such as the interaction between high `predicted_log` values and moderate `mv_traffic` levels, where additional bike lanes led to significant safety improvements. In areas with high traffic, the model indicated that further interventions, like protected bike lanes, would likely be necessary to achieve similar safety benefits.

By aggregating insights from multiple trees, random forest provided a more nuanced understanding of bike lane impacts, complementing the decision tree analysis and offering a more generalizable framework for urban safety planning. However, given the dataset's limitations, its results were not as strong as those from NYC, where the model could leverage a richer and more diverse dataset.

5.3 XGBoost: Advanced Ensemble Learning for Precision

XGBoost was also applied to the San Francisco dataset, building on its strong performance in the NYC analysis, where its advanced ensemble learning capabilities delivered precise predictions. XGBoost's ability to handle complex relationships and interactions made it an appealing choice. However, like Random Forest, XGBoost underperformed in the San Francisco analysis. This can be attributed to the smaller dataset and the challenges of generalizing from a limited number of data points.

As with the NYC data, XGBoost builds decision trees iteratively, with each tree correcting errors from the previous one. Regularization was used to prevent overfitting, which was particularly important given the small dataset. Hyperparameters were tuned using grid search, and early stopping was implemented to optimize performance. The model achieved an impressive Root Mean Squared Error (RMSE) of 25.56, demonstrating high precision. However, this level of precision could be misleading due to the potential overfitting risks associated with the small dataset.

Feature importance analysis highlighted the following key predictors:

- `predicted_log`: With a Gain score of 0.379, this variable was the most influential, capturing the cumulative impact of bike lane development on safety outcomes.
- `mv_traffic`: Scoring 0.323 in Gain, traffic volume played a crucial role in shaping safety outcomes, underscoring the relationship between traffic density and bike lane effectiveness.
- `cumulative_bike_lane`: With a Gain score of 0.298, this variable continued to play an important role in the model, indicating the significant impact of bike lane infrastructure on safety.

XGBoost reinforced insights from both the Decision Tree and Random Forest models, confirming that traffic volume and cumulative bike lane development were key drivers of safety outcomes. The model's advanced ensemble learning approach provided additional precision, validating thresholds and interactions identified in earlier models. However, similar to Random Forest, the model's accuracy was somewhat limited due to the smaller dataset, reducing its ability to generalize effectively.

6 Key Differences and Potential Weaknesses

The San Francisco analysis presents several key differences compared to the New York City (NYC) analysis, which must be considered when interpreting the results:

- **Dataset Size**: The smaller dataset for San Francisco (59 data points) limited model complexity and reduced the ability to uncover intricate patterns. In contrast, NYC's larger dataset (105 data points) allowed for more complex relationships to be identified, making it easier to capture variations in the data.
- **Threshold Stability**: Thresholds identified by Decision Trees, such as the 700,000 vehicles per month split in NYC, may be less stable in the San Francisco data due to its smaller size. In San Francisco, slight changes in the data could lead to shifts in these thresholds, making them less reliable and harder to generalize.

- **Model Performance**: Performance metrics, such as explained variance, were weaker in San Francisco compared to NYC. For example, the Random Forest model for San Francisco explained only 1.87% of the variance, while the NYC model explained 42.59%. This indicates that the San Francisco models were less able to capture the variability in the data.
- **Variable Influence**: The importance of certain predictors, like `predicted_log` and `mv_traffic`, may appear inflated in the San Francisco models due to the smaller dataset, which can lead to overfitting. This reduces confidence in the true influence of these variables on the safety outcomes.
- **Generalizability**: While the NYC models may support broader policy recommendations, the San Francisco models are more context-specific. The smaller dataset and the potential for overfitting make it harder to apply the insights from San Francisco to other regions or broader urban planning strategies.

These differences highlight the challenges of working with smaller datasets and the potential limitations in model reliability and generalizability.

7 RESULTS

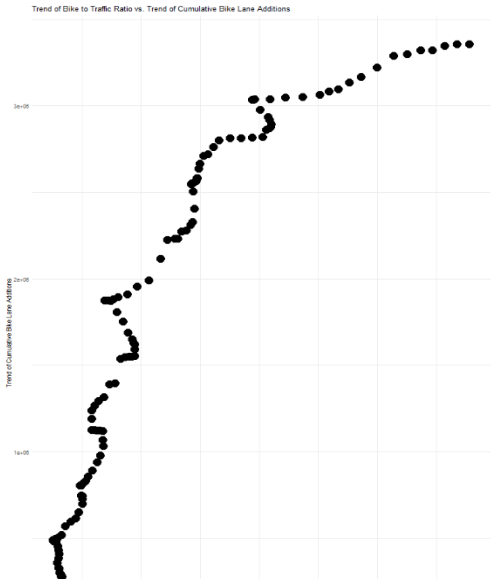


Figure 6; Logarithmic relationship between bike/car ratio and cumulative bike lanes in NYC

7.1 Positive Relationship Between Ratio of Bikes to Cars and Cumulative Bike Lanes

The first and simplest takeaway of the above research is that, after extracting the seasonal component of bike lane usage, the trend of the ratio of bikes to motor vehicle usage is clearly positively correlated with cumulative bike lanes. However, this relationship appears to be logarithmic and not completely linear. Figure 6.

This relationship was observed in both cities studied and appears to be a general rule across cities.

7.2 NYC Accident Rate is Correlated with Various Variables

Both NYC and SF showed similar overall correlations between some key variables in the respective datasets. As noted in the prior section, cumulative bike lanes were positively correlated with the predicted ratio of bike usage to motor vehicle usage, overall greater volumes of traffic were positively correlated with total injuries and deaths and negatively correlated with increased bike lanes. Figure 7.

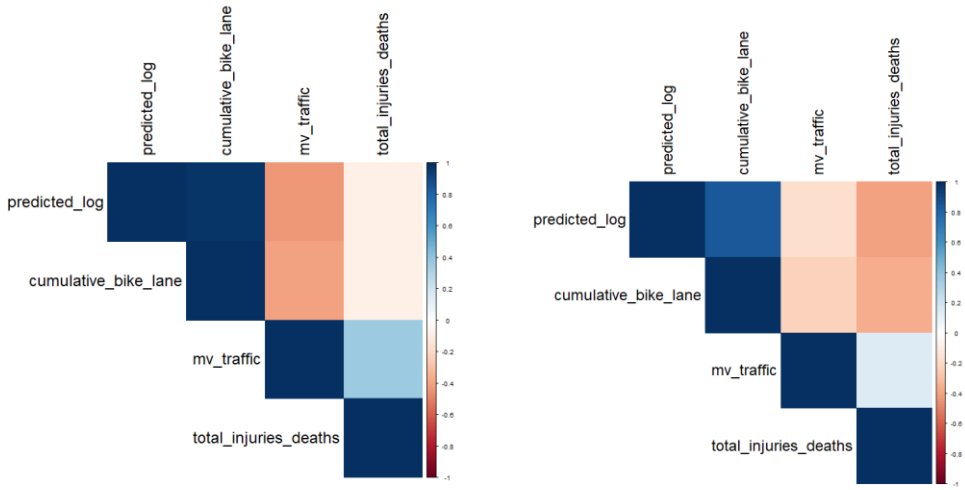


Figure 7; NYC Data Correlation Heatmap, SF Data Correlation Heatmap

7.3 Various Models Tested with Varying Results

Due to the nature of the dataset being both limited in size and duration and the nature of the problem being analyzed, the decision was made to incorporate multiple models in the analysis. Each model provided valuable insight, however, when comparing the multiple models there were different results across the two

cities studied. In NYC the XGBoost model outperformed the other models by a significant margin achieving an RMSE of 0.70 on the scaled NYC dataset. However, the Random Forest model was close enough in performance to warrant inclusion in any final usage of these models due to its easier interpretability. Similarly, the Decision Tree model was kept even with its worse performance due to the nature of public policy often requiring the extreme interpretability provided by the model, Table 2. Other models, such as various kernels SVM models, were tested but failed to provide any reason for inclusion.

Table 2

Metric	Baseline	SVM_RBF	SVR_Sigmoid	DecisionTree	RandomForest	XGBoost
RMSE	1.0594304	0.8658610	0.99166236	0.88735355	0.75156675	0.7031431
MAE	0.7839890	0.5941811	0.74875014	0.64450269	0.53252320	0.5379938
MAPE	0.2891991	-0.1035102	-0.15390186	-0.09098578	0.07133233	-0.1406213
R-squared	0.0000000	0.3029965	0.08574734	0.26796481	0.47486170	0.5403513

Similarly, the San Francisco models also had varying results, with other models shining a bit more. In this dataset, the simple Decision Tree performed the best with the greatest accuracy. The reason for this will be discussed in later sections. However, the Random Forest Model and XGBoost model were retained due to the expectation that with later data those models would outperform in the SF the way they did in NYC, as will be explained later. Table 3.

Table 3

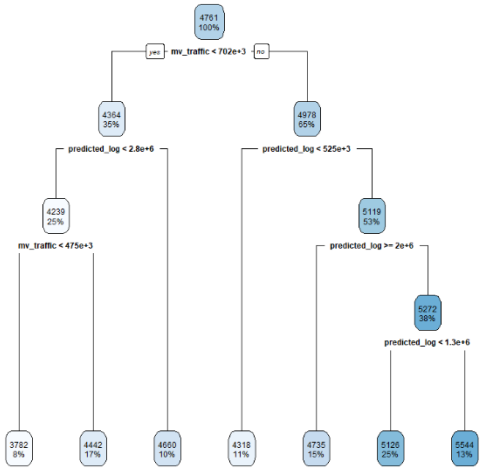
Metric	SVM_RBF	Baseline	XGBoost	DecisionTree	RandomForest
RMSE	0.5874974	0.7561190	0.4287645	0.4180387	0.4926637
MAE	0.5018093	0.6274687	0.3685359	0.3371474	0.4016942
MAPE	-1.0519912	0.1566986	-0.5030096	-0.3939916	-0.5750833
R-squared	0.3414024	0.0000000	0.6492110	0.6665420	0.5368631

7.4 Individual Variable Importance Scores

7.4.1 New York City

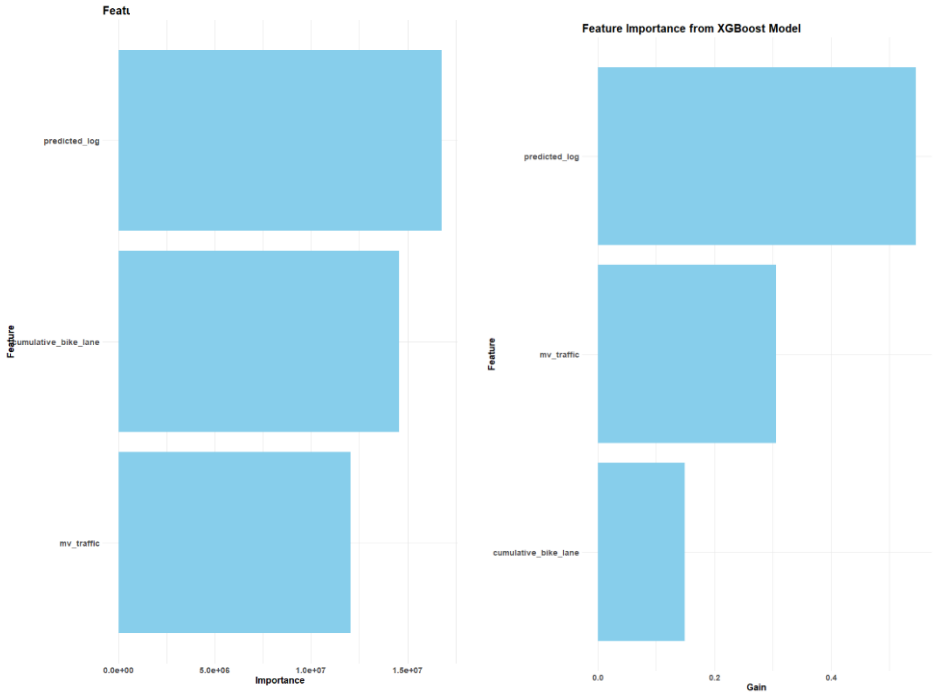
While the results of the three final models vary a bit in the specifics, there was consensus on the importance of specific variables. In the NYC dataset, the single largest indicator of an increase in accidents was the total motor vehicle volume. This was supported by all three final models. The first split of the Decision Tree was at a motor vehicle volume of around 700,000 vehicles. However, the predicted log variable, a measure of the expected ratio of bike to motor vehicle usage, featured the most prominently in the actual tree. The model split on this variable very frequently, indicating a high significance. Figure 8.

Figure 8



Similarly, plotting the feature importance of both the Random forest model and the XGBoost models yields similar results. In both models the predicted_log variable was clearly the most important feature, indicating that an increase in the ratio of users opting for bikes will be the biggest indicator of a decrease in pedestrian accidents. The models differed however, in regards to the order of the importance of the other two variables. The Random Forest model showed cumulative bike lanes as the next biggest factor while the XGBoost model resulted the motor vehicle traffic having the second largest impact. It is important to note the results of the prior section 7.1 indicating a positive relationship between cumulative bike anes and the ratio of bike to motor vehicle users. This gives cumulative bike lanes an indiriect correlation with the total pedesterian accidents and can be the cause of the difference between the models. Figure 9.

Figure 9; Feature Importance for NYC Random Forest and XGBoost Models



7.4.2 San Francisco

At first glance, the results of the San Francisco models appear to be very different than the NYC models. The Decision Tree only contains one split and completely ignores the motor vehicle variable. However, upon further reflection, the models of the two cities are perfectly in sync. San Francisco does not have the population density that NYC has, nor the overall total motor vehicle volume. This effectively negates the first split of the decision tree and leaves it with the variable identified in the NYC models as the most important feature in predicting pedestrian accidents, predicted log, or the expected ratio of bike users to car users. Figure 10.

Figure 10; SF Decision Tree

As with the NYC models, the Random Forest and XGBoost feature importances that reinforce these findings. The predicted ratio variable stands out as the clearest

indicator of pedestrian accidents in both models, fitting the findings of both the NYC models and the SF Decision Tree model. As with the NYC findings, the Random Forest and XGBoost models disagree about the order of the second and third most important features, with the Random Forest valuing cumulative bike lanes as the second most important followed by motor vehicle traffic and the XGBoost model having that order reversed. Figure 11.

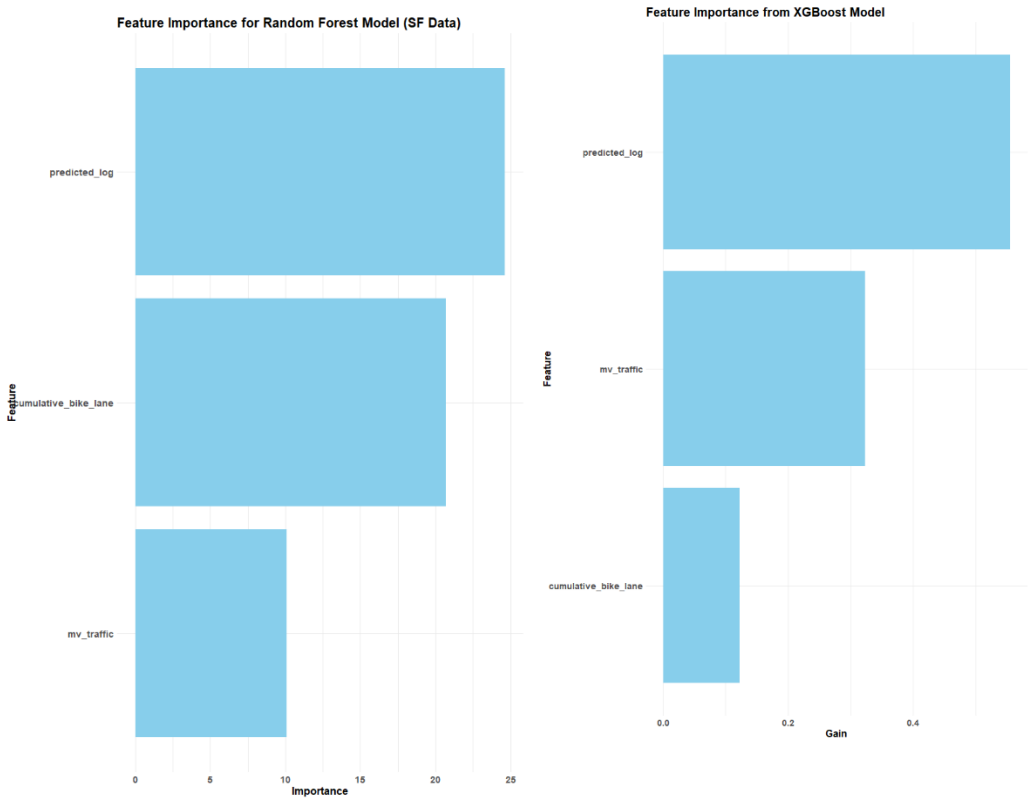


Figure 11; Feature Importance of the SF Random Forest and XGBoost Models

7. DECISION MAKERS: TRANSLATING ANALYSIS INTO ACTIONABLE URBAN PLANNING INSIGHTS

As cities prioritize cycling infrastructure, decision-makers face challenges in ensuring safety and maximizing the impact of bike lane investments. Findings from this research, derived from Decision Trees, Random Forest, and XGBoost

models, provide a robust framework to guide decisions, enabling policymakers to translate data into actionable insights for improving urban safety.

7.1 Cumulative Bike Lane Development Positively Influences Safety and Usage

Both the NYC and SF analyses revealed a clear positive logarithmic relationship between cumulative bike lane development (predicted_log) and the trend of the ratio of bikers to motor vehicle users. This relationship underscores the critical role of bike lanes in promoting a modal shift toward cycling. Increasing cumulative bike lane infrastructure not only improves safety outcomes but also encourages more people to switch to cycling, creating a virtuous cycle of safer streets and greater bike usage.

7.2 Motor Vehicle Volume is the Key Factor for Pedestrian Safety

The NYC dataset highlighted that motor vehicle volume is the most significant factor impacting pedestrian safety. In areas or time periods with high traffic volumes, the effectiveness of bike lane additions was mitigated. This suggests that resources in such settings are better allocated to complementary measures, such as traffic calming initiatives, protected bike lanes, or enhanced pedestrian infrastructure. While the SF dataset did not include motor vehicle volume as a key variable in its final decision tree, this is likely because SF lacks the extreme traffic volumes observed in NYC. This distinction highlights the importance of local context when planning safety interventions.

7.3 Ratio of Bike to Car Usage (Predicted_Log) is Crucial for Safety Insights

The predicted logarithmic ratio of bike to car usage was identified as a critical variable for understanding accident trends in both datasets. In NYC, it was the second split in the decision tree, clearly influencing safety outcomes. For SF, it was the sole variable used in the model, reflecting its strong predictive power in that dataset. The close relationship between cumulative bike lane development and the predicted ratio further emphasizes the interconnected benefits of investing in bike infrastructure.

7.4 High-Traffic Areas Require Additional Measures

Both datasets consistently showed that bike lanes alone are insufficient in high-traffic areas. In these environments, additional interventions, such as protected bike lanes, dedicated bike signals, or speed reductions, are necessary to achieve meaningful safety improvements. This reinforces the need for a layered, context-specific approach to urban safety planning.

7.5 Identifying Thresholds for Effective Interventions

Clear thresholds emerged in both datasets that delineate where the effectiveness of bike lane interventions begins to diminish. For example, in NYC, bike lanes became less effective in areas with more than 35,000–40,000 vehicles per month. These thresholds provide valuable guidance for urban planners, enabling them to prioritize resources and focus on areas with the greatest potential for impact.

7.6 Thresholds and Variables Vary Across Cities

While the key variables affecting safety and bike usage are consistent across cities, their specific thresholds can vary significantly. For instance, SF and NYC showed different splitting thresholds for traffic volume and bike lane development. This underscores the importance of tailoring interventions to the unique characteristics of each city. New cities should ideally be evaluated using similar model training procedures to identify their optimal thresholds.

7.7 Data-Driven Models Inform Targeted Urban Planning

The analyses demonstrated the power of data-driven decision-making in urban planning. By leveraging models such as Decision Trees, Random Forest, and XGBoost, policymakers can identify actionable insights, such as the critical role of cumulative bike lane development and traffic volume. These models allow cities to optimize their resources, prioritize effective interventions, and make informed decisions that improve cyclist and pedestrian safety.

8. DISCUSSION

8.1 Increased Bike Usage: A Shift Towards Sustainable Transportation

The findings from this study highlight a clear correlation between the addition of bike lanes and a marked increase in bike usage across cities. This aligns with the research by Kraus and Koch (2021), which observed substantial increases in cycling following the installation of bike lanes. In our analysis, we also observed that dedicated cycling infrastructure significantly encourages more people to choose biking as a primary mode of transportation, a trend especially noticeable in cities with robust bike-share systems like Citi Bike in New York and Bay Wheels in San Francisco. These bike-share systems amplify the positive effects of infrastructure investments by offering more people access to bikes, making cycling a more convenient and viable transportation option.

The evidence suggests that bike lanes serve as a catalyst for broader shifts in transportation behavior. By creating safer and more accessible spaces for cyclists, cities can help reduce reliance on motor vehicles, which has multiple downstream benefits. For example, cities with more bike lanes experience lower traffic congestion, improved air quality, and healthier communities, as biking provides an alternative to car usage.

Our findings also support this view, with specific numbers indicating that the increase in bike usage is not merely a seasonal trend but a lasting shift. For example, in New York City, the presence of additional bike lanes correlates with a substantial increase in biking, as reflected in the model's statistics—such as the decrease in `mv_traffic` and corresponding increase in bike ridership. This positive relationship did not appear to be linear though, the overall relationship best fits a logarithmic relationship. The earliest additions of bike lane infrastructure pay off the most in terms of increasing bike usage. As the network grows, the overall returns tend to diminish.

8.2 Decreased Car Dependency: Relieving Urban Congestion

A related and highly dependent benefit of the aforementioned increase in bike usage is the reduction in car dependency associated with the construction of new

bike lanes. As bike usage increases, car ridership either decreases or grows at a slower rate, depending on the city. This result corroborates the hypothesis that bike lanes can help alleviate congestion by redistributing travel demand from cars to bicycles. The models from New York and San Francisco show that bike lane additions were often accompanied by a noticeable decline in car usage, particularly in areas where bike lanes were implemented extensively. For instance, in New York, the regression model shows that as bike lane miles increased, the `mv_traffic` variable consistently dropped, with key thresholds showing significant shifts in travel patterns.

Hartgen and Fields' (2009) argument that reducing traffic congestion can boost economic productivity by improving travel efficiency is particularly relevant. While our study does not measure productivity directly, we observe that reduced car dependency contributes to better mobility and less time lost in traffic. Furthermore, the observed reduction in traffic aligns with findings from other studies, where decreased car usage was found to correlate with economic benefits. In cities like Chicago, for example, areas with extensive bike lane networks showed increased pedestrian and foot traffic, a trend that could be indicative of a shift toward more sustainable local economies.

8.3 Enhanced Pedestrian Safety: Creating Safer Streets

A major contribution of this research is its exploration of pedestrian safety, an area that has been relatively underexplored in existing literature on bike lanes. Our study reveals a noticeable decrease in pedestrian accidents in areas where new bike lanes were added, particularly when the bike lanes were designed with protective barriers. This finding echoes the work of Buehler and Dill (2015), who concluded that dedicated cycling spaces improve traffic organization and reduce pedestrian injuries. Moreover, our analysis further underscores that the design and placement of these bike lanes are crucial in determining their impact on safety. For instance, protected bike lanes—those separated from both vehicular and pedestrian traffic—were associated with the greatest reduction in pedestrian accidents.

The findings present actionable insights for city planners. By designing bike lanes with safety in mind, cities can create safer streets for all road users, not just cyclists. Furthermore, the reduction in pedestrian accidents does not just save lives but also boosts public confidence in alternative modes of transportation, encouraging more people to walk or bike rather than drive. Additionally, our data highlights an interesting seasonal pattern in bike usage, with peaks during the warmer months. This seasonal trend also seems to influence the timing of bike lane expansions, with cities often adding infrastructure during construction-friendly seasons. This seasonal dynamic emphasizes the importance of strategic planning to ensure that bike lane expansions are rolled out in conjunction with periods of high demand, optimizing their immediate impact.

8.4 Economic and Spatial Considerations

The financial and spatial challenges associated with bike lane construction remain important considerations. While the initial costs, potential disruption to traffic, and the reduction in parking spaces may present barriers, the long-term benefits—such as increased bike usage, reduced car dependency, and improved pedestrian safety—justify these investments. As noted by Arancibia et al. (2019), bike lanes can stimulate local economic activity by attracting more foot and bike traffic, suggesting that their benefits extend beyond transportation improvements to broader urban vitality.

Our study's results are consistent with this perspective. For instance, in New York, areas with new bike lanes saw a marked increase in pedestrian traffic, which has been shown to benefit local businesses. Similarly, in San Francisco, the combination of bike lanes and bike-share stations led to higher ridership and improved local economic outcomes, demonstrating that bike lanes are not just a transportation infrastructure investment, but an economic one as well.

8.5 Statistical Insights

The statistical models employed in this study, particularly the Random Forest and XGBoost models, provide robust evidence of the relationship between bike lane expansion and shifts in transportation behavior. A key predictor in these models

is predicted_log, which represents the predicted trend in the ratio of bike to motor vehicle users as a function of cumulative bike lane expansions. This variable captures how changes in infrastructure can influence the relative use of bikes versus cars over time.

For example, in New York, the Random Forest model indicates that predicted_log has an importance score of 16,759,167, highlighting its significant role in explaining the variation in bike usage. This means that as bike lane mileage increases, the model predicts a substantial shift towards biking relative to motor vehicle use. Similarly, the XGBoost model for both New York and San Francisco further supports this trend, with predicted_log consistently emerging as a key predictor of bike ridership increases.

These findings suggest that the ratio of bike to motor vehicle usage is not only positively correlated with bike lane expansions, but it also indicates the critical role that bike infrastructure plays in shifting transportation patterns in favor of sustainable mobility. The predicted_log variable thus serves as a powerful tool for understanding and forecasting the impact of bike lane investments, providing urban planners with actionable insights for future infrastructure planning.

8.6 Generalization to Other Cities

While this study focused on the impact of bike lane infrastructure in New York City and San Francisco, the methodology and key variables are highly transferable to other urban environments. The core variables utilized in the analysis, such as bike lane mileage, traffic volume, bike usage, pedestrian accidents, and population data, are relevant in nearly any city with available data. Given that these variables are common across many urban centers, the models developed for New York and San Francisco can serve as a useful framework for understanding the potential impacts of bike lane expansion in other cities.

8.7 Model Transferability and Threshold Identification

One of the strengths of this approach is its flexibility. By utilizing similar features and models, the same analytical framework can be applied to other cities. The Random Forest and XGBoost models used in this study have shown their ability

to capture the complex relationships between bike lane infrastructure, traffic patterns, and safety outcomes. Therefore, the same approach can be applied to other cities by simply substituting the local data for the existing predictors.

To adapt the models to a new city, key steps would involve:

1. Data Collection: Gather similar datasets to those used in New York and San Francisco.
2. Model Training: Run the city's data through the same Random Forest or XGBoost models.
3. Threshold Identification: Once the model is trained, identify key thresholds for bike lane additions.
4. Evaluation: Evaluate the models by comparing predicted outcomes against actual data.

8.8 Expected Outcomes

Given the robust nature of the models developed for NYC and SF, we would expect similar results when applying these models to other cities with comparable infrastructure and data. For instance, cities like Chicago, Portland, and Seattle should see similar patterns: increased bike usage, reduced car traffic, and improved pedestrian safety in areas with expanded bike infrastructure.

9 LIMITATIONS

This study has several limitations that should be considered when interpreting the findings. First, the dataset used for analysis is limited both in scope and duration. The primary data on bike lane infrastructure and Citi Bike usage in New York City (NYC) and San Francisco (SF) spans only a relatively short period, which may not fully capture long-term trends or the broader effects of infrastructure changes on cycling behavior or accident rates. As bike lane infrastructure continues to evolve, future studies may need to account for the dynamic nature of urban transportation systems, which could alter the model's predictive power over time.

Second, the analysis is geographically constrained to NYC and SF. While these cities represent two major urban areas with differing cycling cultures and

infrastructure, the findings may not be directly applicable to other cities. Although we are hopeful that the model could be adapted to other major cities, this remains unproven. Cities with different sizes, traffic patterns, or cycling cultures may experience different results, and factors such as local policy changes or variations in urban density could significantly impact the applicability of the model. Additionally, the scale of the cities in this study differs, with NYC being a much larger metropolitan area compared to SF. The vast size and complexity of NYC introduces a range of unaccounted-for variables—such as socioeconomic factors, traffic congestion, and varying levels of enforcement of traffic laws—that contribute to the city's accident rates. Consequently, even a well-calibrated model with high correlation may only account for a small fraction of the variability in accident rates, particularly in a city as large and complex as NYC.

Furthermore, the data used in this study were sourced from disparate origins, which introduces potential issues with data integrity and consistency. For example, data on bike lane infrastructure and Citi Bike usage were collected from different agencies and sources, each with its own set of standards and reporting methods. This variability in data quality and measurement could affect the accuracy and reliability of the results, particularly when combining datasets for modeling purposes.

Lastly, while this study focused on the impact of bike lane infrastructure and bike-share usage on accident rates, there are numerous other factors that could influence these outcomes, which were not captured in this analysis. These include factors such as the quality of road maintenance, the effectiveness of traffic laws and enforcement, weather conditions, demographic factors, and the overall prevalence of other transportation modes. The inability to include these variables limits the comprehensiveness of the model and may lead to an oversimplification of the complex relationship between infrastructure and safety outcomes.

In conclusion, while this study provides valuable insights, the aforementioned limitations highlight the need for further research and more granular data collection in order to build more robust models that can be generalized across cities with varying conditions and over longer time periods.

10 CONCLUSION

The results from the analyses suggest a positive relationship between bike lane infrastructure and increased bike usage, with this relationship appearing to follow a logarithmic pattern, showing diminishing returns at larger scales. Additionally, a statistically significant association was found between a higher proportion of transit users opting for bikes instead of motor vehicles and pedestrian accident rates, particularly at lower traffic volumes (around 700,000 vehicles per month in NYC). While the exact threshold may vary across cities—due to factors such as population density, weather, and traffic count methods—the findings from New York City and San Francisco indicate that the framework developed in this study has potential for broader application in cities with similar infrastructure and conditions.

By using a consistent set of variables and modeling techniques, urban planners can apply this framework to assess the impact of bike lane expansions in their own cities, identify key thresholds for significant shifts in transportation patterns, and make informed decisions about infrastructure investments. This approach is particularly useful not only for cities with well-established bike networks but also for those starting to explore the potential benefits of bike lanes.

In conclusion, the analytical tools presented in this study provide a scalable framework to support sustainable urban mobility and safer streets. By embracing such tools, cities can make data-driven decisions that promote healthier, more sustainable transportation options and ultimately contribute to safer urban environments for all residents.

Appendix

The data for this project was sourced from various locations and compiled into distinct datasets for each city, containing all relevant information. These datasets, along with the code used to create, clean and wrangle them can all be found in the github repository linked below. The primary data sources include some extremely large files and are not in that repository.

The code to generate these models, including some attempts not included in the above writeup, are all included in this repository too, in a file named Final_Draft.rmd. This file contains the code generating the above visualizations as well, including both the charts and tables.

Shayaeng/Data698 (github.com)

Data Sources

NYC Accident Data - https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu/about_data

NYC Bike Count Data - https://data.cityofnewyork.us/Transportation/Bicycle-Counts/uczf-rk3c/about_data

Citi Bike Data - https://citibikenyc.com/system-data/operating-reports

NYC Bike Route Data - https://data.cityofnewyork.us/dataset/New-York-City-Bike-Routes/mzxg-pwib/about_data

NYC Traffic Volume - https://data.cityofnewyork.us/Transportation/Automated-Traffic-Volume-Counts/7ym2-wayt/about_data

SF Bike Route Data - https://data.sfgov.org/Transportation/MTA-Bike-Network-Linear-Features/ygmz-vaxd/about_data

SF Pedestrian Accident Data - https://data.sfgov.org/Public-Safety/Traffic-Crashes-Resulting-in-Injury/ubvf-ztfx/about_data

Bay Wheels Data - https://s3.amazonaws.com/baywheels-data/index.html

SF Traffic Data - https://www.sfmta.com/TrafficCounts

References

1. **Combs, T. S., & Pardo, C. F. (2021)**. Shifting streets COVID-19 mobility data: Findings from a global dataset and a research agenda fo

r transport planning and policy. Transportation Research Interdisciplinar y Perspectives, 9, 100322. https://doi.org/10.1016/j.trip.2021.100322

2. **Rérat, P., Haldimann, L., & Widmer, H. (2022)**. Cycling in the era of Covi d-19: The effects of the pandemic and pop-up cycle lanes on cycling practices. Transportation Research Interdisciplin ary Perspectives, 15, 100677. https://doi.org/10.1016/j.trip.2022.100677
3. **Karpinski, E. (2021)**. Estimating the effect of protected bike lanes on bike-share ridership in Boston: A case study on Commonwealth Avenue. Case Studies on Transport Policy, 9(3), 1313-1323. https://doi.org/10.1016/j.cstp.2021.06.015
4. **Buck, D., & Buehler, R. (2011)**. Bike Lanes and Other Determinants of Ca pital Bikeshare Trips. TRB 2012 Annual Meeting Paper. Virginia Tech Ale xandria Center.
5. **Prud'homme, R., & Lee, C.-W. (1999)**. Size, Sprawl, Speed and the Efficiency of Cities. Urban Studies, 36(11), 1849-1858. http://www.jstor.org/stable/43198143
6. **Hartgen, D. T., & Fields, M. G. (2009)**. Gridlock and Growth: The Effect of Traffic Congestion on Regional Economic Performance. Reason Foundation.
7. **Kraus, S., & Koch, N. (2021).** Provisional COVID-19 infrastructure induces large, rapid increases in cycling. Proceedings of the National Academy of Sciences, 118(15), e2024399118. https://doi.org/10.1073/pnas.202439911
8. **Gatera, A., Kuradusenge, M., Bajpai, G., Mikeka, C., & Shrivastava, S. (2023).** Comparison of random forest and support vector machine regression models for forecasting road accidents. *Scientific African*, 15, e01739. https://doi.org/10.1016/j.sciaf.2023.e01739