Final Project

Analyzing Climate's Role in Accidents in Nashville, U.S.A.

Sergio David Escobar

Data Scientist

Abstract

This project explores the relationship between climate conditions and the occurrence of accidents in Nashville, U.S.A., utilizing various data science methodologies including exploratory data analysis, statistical testing, and predictive modeling.

October 4, 2023

Contents

1 Introduction

1.1 Objective

The primary objective of this project, titled "Analyzing Climate's Role in Accidents in Nashville, U.S.A.," is to explore and quantify the relationship between various weather conditions and the frequency of traffic accidents in selected U.S. cities. With a principal focus on Nashville and secondary analyses on Madison and Boise, this project aims to understand how different weather variables influence the incidence of accidents. By employing various statistical analyses, data exploration techniques, and predictive modeling, this project seeks to derive insights, uncover patterns, and forecast accident occurrences, contributing to the development of informed safety measures and traffic management solutions in different weather conditions.

1.2 Overview

The methodology adopted in this project encompasses a structured approach involving extensive data exploration, cleaning, transformation, and exploratory data analysis (EDA) to understand the underlying patterns and relationships in the data. Time-series analysis techniques are deployed to examine seasonal trends and patterns in accident occurrences over time. Statistical tests are conducted to validate the significance of the observed trends and relationships between weather variables and accident incidences.

Several predictive models, including SARIMA, Random Forest, XGBoost, LightGBM, and LSTM, are built and evaluated to forecast "Total Accidents" based on observed patterns and weather conditions. These models are chosen due to their versatility, capability to handle multiple input features, and effectiveness in modeling complex relationships within the data. The performance of each model is assessed using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

1.3 Access to Additional Resources

For a comprehensive view of all graphs, images, and detailed code used in this analysis, please refer to the accompanying Jupyter Notebook available in the project's GitHub repository.

2 Dataset Overview

The primary dataset used in this study was a countrywide car accident dataset, which covers 49 states of the USA. The data were collected from February 2016 to March 2023. This dataset is crucial for our analysis as it offers a comprehensive overview of accident occurrences and their corresponding conditions across a wide temporal and spatial scope.

Additionally, weather data essential for the project were sourced from the Meteostat API, which employs several Python libraries to ensure the accuracy and comprehensiveness of the information (3).

2.1 Data Collection

The accident data were meticulously collected, cleaned, and transformed to remove inconsistencies and provide additional variables for further insights (1). Similarly, the weather data collection involved downloading raw data and undergoing a series of cleaning and transformation processes to append additional variables and remove inconsistencies (3).

2.2 Description

The dataset utilized in this project comprises diverse features offering insights into weather conditions, geographical information, and timestamps of accidents in Nashville, Madison, and Boise. Features include:

Season: Categorical variable indicating the meteorological season.

Temperature(C): Continuous variable representing the temperature in Celsius.

Visibility(km): Continuous variable depicting visibility in kilometers.

Weather Condition: Categorical variable describing the prevailing weather condition.

Precipitation(mm): Continuous variable representing the precipitation in millimeters.

Total Accidents: Continuous variable indicating the total number of accidents recorded.

Each entry in the dataset represents a daily record, with weather variables averaged, summed, or mode-selected as appropriate, and "Total Accidents" summed for the day.

2.3 Source

The primary dataset utilized is the US-Accidents dataset, aggregated from multiple streaming traffic incident data APIs (1). This dataset is extensively used to study various aspects of road accidents. The weather condition data was sourced from the Meteostat API, which emphasizes accuracy and comprehensiveness in providing weather-related information (3).

3 Data Cleaning, Preprocessing, and Transformation

3.1 Initial Exploration

The first step in our analytical process involved an initial exploration of the dataset to understand its structure, features, and statistical properties. I examined individual datasets for each city—Nashville, Madison, and Boise—focusing primarily on Nashville for an in-depth analysis. The exploration involved understanding the data types of each feature, identifying the presence of any missing values, and observing the basic statistical properties such as mean, median, standard deviation, minimum, and maximum values of the variables.

This exploration revealed a combination of categorical and continuous variables, with features like "Season" and "Weather Condition" being categorical, and "Temperature(C)", "Visibility(km)", "Precipitation(mm)", and "Total Accidents" being continuous. A summary of the statistical properties provided insights into the central tendency, dispersion, and distribution of the continuous variables, while the frequency distribution of categorical variables was examined to understand the diversity and prevalence of different categories within those features.

3.2 Data Cleaning and Preprocessing

The assurance of data quality was of utmost importance. The cleaning phase tackled missing values, outliers, and data type consistency.

Handling Missing Values: An examination of the dataset revealed missing values in specific columns:

- Wind_Direction: Contained 153 missing values.
- Weather_Condition: Included 87 missing entries.
- Street: Presented 39 missing data points.

For categorical features like 'Wind_Direction' and 'Weather_Condition', missing values were imputed either with the mode of the column or labeled as 'Unknown'. For the

'Street' feature, missing entries were replaced with the 'Unknown' placeholder. This intervention rendered the dataset free from missing values.

Addressing Outliers: Outliers, due to their potential to skew results, were addressed meticulously. The dataset, which originally spanned 72,670 rows, was pruned to 44,650 rows post outlier removal using the IQR method.

Data Type Conversions and Feature Engineering: Ensuring that each feature was represented using the right data type was pivotal. Specifically, the 'Start_Time', 'End_Time', and 'Weather_Timestamp' columns underwent transformation into datetime objects, facilitating subsequent analyses. Moreover, the 'Duration' feature was introduced to capture the accident's span in minutes, computed from the 'Start_Time' and 'End_Time' columns.

Insights Derived: The data cleaning and preprocessing phases fortified the dataset's integrity, making it apt for subsequent analysis, statistical validation, and predictive modeling. Through these rigorous procedures, the dataset was honed for completeness, consistency, and relevancy, providing a solid foundation for the next stages of the project.

3.3 Transformation for Analysis and Modeling

The dataset underwent significant transformation to suit the requirements of machine learning models. The engineering encompassed the introduction of new features and preprocessing for specific models.

Feature Engineering: Recognizing the time-series nature of the data, additional features were crafted. Lagged values of "Total Accidents" over seven days, echoing the weekly seasonality, were introduced. This addition helped models in capturing past trends, thereby understanding the inherent temporal dependencies.

Scaling: To ensure consistency and numerical stability, the data was scaled using MinMaxScaler.

Preparing Sequences: The time-series data was transformed into sequences with a fixed length of seven days to reflect the weekly seasonality, making it apt for models that require sequential input.

These meticulous steps in data transformation were instrumental in ensuring that the dataset was primed for exploratory data analysis, statistical validation, and predictive modeling.

4 Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

A comprehensive univariate analysis was undertaken to decipher the distribution, central tendencies, and dispersion of individual attributes within the dataset. This form of preliminary exploration employed histograms for numerical attributes and bar plots for categorical ones.

Numerical Variables: The distributions of numerical features, including `Severity`, `Temperature(C)`, `Visibility(km)`, `Wind_Speed(kmh)`, and `Duration`, were visualized using histograms. The analysis revealed:

- The majority of accidents are categorized under severity level 2.
- The temperature distribution is approximately normal, predominantly centered around 15°C to 20°C.
- The visibility distribution leans towards higher values, indicating that most accidents occur under clear visibility conditions.
- Wind speed demonstrates a rightward skew, suggesting that a larger proportion of accidents happen at lower wind speeds.
- The distribution of accident duration is heavily skewed to the right, indicating that most accidents are resolved quickly, with only a few taking a longer time.

Categorical Variables: For categorical variables, specifically `Weather Condition` and `Sunrise Sunset`, bar plots were employed. The visualizations indicated:

- Most accidents occur during "Clear" weather conditions, followed by "Overcast" and "Mostly Cloudy", implying that a significant number of accidents take place under relatively good weather conditions.
- Accidents are more frequent during the day as opposed to the night.

Insights Derived: The univariate analysis enriched the understanding of each variable's inherent characteristics, offering insights into their distributions, the prevalence of specific conditions, and the general trend within the dataset.

4.2 Bivariate Analysis

The bivariate analysis was designed to probe the relationships between diverse meteorological variables and accident incidences. This incorporated scatter plots and correlation matrices to fathom the interdependencies between the variables.

Scatter Plots: Scatter plots were crafted to visualize the relationships between `Temperature(C)`, `Visibility(km)`, `Wind_Speed(kmh)`, and `Severity`. However, these plots insinuated a lack of a conspicuous linear relationship between these meteorological variables and accident severity.

Correlation Matrices: The correlation matrix, constructed to quantify the relationships between different attributes, resonated with the scatter plot observations. No discernible correlations emerged between `Severity` and other variables like `Temperature(C)`, `Visibility(km)`, and `Wind_Speed(kmh)`. It was also noted that the correlation coefficients for `Severity` were `NaN`, possibly stemming from the discrete nature of the variable, hinting at the need for alternative analytical methods.

Insights Derived: Bivariate analysis shone light on the interrelations between meteorological factors and accident occurrences. Though clear linear relationships weren't visible, the analysis underscored the importance of embracing more complex or non-linear models to comprehend the intricate relationships embedded within the dataset.

4.3 Time-Series Analysis

The time-series nature of the accident data necessitated a comprehensive analysis to illuminate underlying patterns, trends, and seasonalities. This entailed examining the number of accidents over time and implementing seasonal decomposition to discern any inherent patterns.

Temporal Progression of Accidents: By aggregating the dataset to a daily frequency, the number of accidents in Nashville over time was visualized. The resulting graph showcased evident fluctuations, with certain days marking a more significant number of accidents than others.

Seasonal Decomposition: To further elucidate the inherent structures of the time series, seasonal decomposition was applied. This method unraveled the time series into its constituent components:

- **Trend Component:** Representing the long-term progression in the data, this component abstracts from daily fluctuations to spotlight the overarching trajectory of accident occurrences.
- **Seasonal Component:** Highlighting the recurring short-term cycles in the dataset, this component revealed a clear weekly seasonality in the number of accidents.
- **Residual Component:** Capturing the error or the noise after extracting the trend and seasonal components, this residual analysis can sometimes unearth insights into atypical events or anomalies.

Insights Derived: The time-series analysis was instrumental in demystifying inherent temporal patterns, discerning the seasonal variations, and identifying long-term trends within the accident data. This deepened comprehension of the time-dependent phenomena vital for crafting accurate and effective time-series forecasting models.

Through a detailed exploration, the analysis unearthed invaluable insights about the individual characteristics of the time series, its intrinsic temporal structures, and its periodic fluctuations. This robust analysis sets a firm groundwork for the upcoming phases of statistical validation and predictive modeling.

5 Statistical Tests

5.1 Tests Performed

The significance of observed trends and relationships identified during the Exploratory Data Analysis (EDA) was validated using specific statistical tests. The main intention was to determine if the observed patterns and relationships between weather variables and "Total Accidents" were statistically significant or if they arose by mere chance.

Test for Stationarity (Augmented Dickey-Fuller Test): This test was executed to assess whether the time series data for "Total Accidents" is stationary, meaning it does not have time-dependent structures like trends or seasonality.

Test for Seasonality (Autocorrelation and Partial Autocorrelation): The autocorrelation function (ACF) and partial autocorrelation function (PACF) were analyzed to ascertain and quantify the seasonality present in the "Total Accidents" time series data.

5.2 Results

Augmented Dickey-Fuller Test: The ADF statistic returned a value of -3.4163 with a p-value of 0.0104. Given that the p-value is less than the commonly used significance level of 0.05, it implies the rejection of the null hypothesis, confirming that the time series data for "Total Accidents" is stationary.

Autocorrelation and Partial Autocorrelation Analysis: The ACF plot indicated significant autocorrelation at multiple lags, notably at lags 7, 14, and 21. This suggests a weekly seasonality in "Total Accidents". The PACF plot further confirmed this weekly seasonality by showing significant partial autocorrelation at the same lags.

5.3 Conclusions Drawn

The statistical tests played a pivotal role in validating the insights gathered from the EDA. The Augmented Dickey-Fuller test confirmed the stationarity of the "Total Accidents" time series, ensuring that it meets an essential assumption for various time series forecasting models. The ACF and PACF plots quantified the weekly seasonality in the data, offering valuable insights for subsequent modeling.

By establishing the stationarity of the dataset and quantifying the observed seasonality, these tests solidified the foundation for the subsequent modeling phase. They ensured that the analysis and modeling are rooted in statistically validated insights, heightening the reliability of the final conclusions and recommendations derived from the project.

6 Model Building and Evaluation

6.1 Model Selection

Several models were selected to forecast "Total Accidents" based on observed patterns and weather conditions. The models were chosen due to their ability to work with time-series data, accommodate multiple input features, and effectively capture intricate relationships and patterns within the data. The selected models include:

SARIMA: Chosen for its competence in modeling time-series data with a seasonal component. It inherently manages trends and seasonality in the data, rendering it suitable for forecasting "Total Accidents".

Random Forest: An adaptable ensemble learning method, it was picked for its capacity to manage high-dimensional data and its robustness against overfitting, offering insights into feature significance.

XGBoost and LightGBM: These gradient boosting algorithms were opted for their efficiency in modeling intricate relationships, handling varied data types, and their capabilities in feature selection.

LSTM: Recognized for its expertise in learning long-term dependencies in sequence data, LSTM was chosen to grasp the sequential patterns in time-series data, offering the skill to model complex temporal relationships.

6.2 Feature Selection

The features used for model building comprise weather variables and lagged values of the target variable, including "Temperature(C)", "Visibility(km)", "Precipitation(mm)", "Season", "Weather Condition", and lagged values of "Total Accidents". Incorporating lagged values allows models to learn from past observations, hence understanding temporal dependencies in the data.

6.3 Model Training

Each model underwent a thorough training process:

SARIMA: Configured to manage seasonality and trends, its parameters were optimized based on the AIC criterion, ensuring the best fit to the training data.

Random Forest, XGBoost, and LightGBM: Trained using a mix of weather variables and lagged values of "Total Accidents". Hyperparameter tuning was executed using grid search, focusing on minimizing prediction errors.

LSTM: Trained using sequences of past observations, it learned long-term dependencies and temporal patterns in the data. Training included adjusting sequence length and model parameters for optimal performance.

6.4 Model Evaluation

The models were assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics offered insights into the accuracy and generalizability of the models on the testing set. The performance of each model was reviewed to gauge its capability to accurately predict "Total Accidents" based on observed patterns and weather conditions.

6.5 Model Comparison

Based on the results:

Model	MAE	RMSE	MAPE
SARIMA	14.51	3.81	69.72%
Random Forest	10.40	12.40	47.89%
XGBoost	10.52	12.28	44.15%
LightGBM	10.37	12.64	43.52%
LSTM	0.09 (scaled)	0.13 (scaled)	84.60%

The models showcased varied capabilities in capturing the underlying patterns and predicting "Total Accidents". The performance metrics highlighted strengths and limitations inherent in each model, giving a clear picture of their predictive abilities.

7 Insights, Conclusions, and Recommendations

7.1 Insights

Delving deep into the dataset, especially with a focus on the city of Nashville, brought forth several pivotal insights:

Impact of Weather Conditions: While the Univariate and Bivariate Analyses spotlighted the influence of specific weather conditions on the number of accidents, an essential observation was the higher accident rates during seemingly benign weather conditions like clear skies. This might be attributed to the predominance of such weather conditions. Variables such as "Temperature(C)", "Visibility(km)", and "Wind_Speed(kmh)", though explored, did not manifest a pronounced linear relationship with accident occurrences.

Presence of Seasonality: The data, when subjected to Time-Series Analysis and Seasonal Decomposition, revealed unmistakable seasonal patterns. A standout observation was the weekly seasonality, hinting at certain days being more susceptible to accidents.

Modeling Insights: The journey of predictive modeling was scaffolded by data transformations and adept feature engineering. A spectrum of models was tested, ranging from traditional time-series models to machine learning paradigms like Random Forest. The LSTM model, despite its intricate capabilities in handling time-dependent data, was overshadowed by models like LightGBM in this dataset, particularly when judged on the yardstick of MAPE.

7.2 Conclusions

The meticulous exploration and scrutiny of the dataset culminated in the understanding that weather conditions, especially in Nashville, bear a tangible imprint on the number of traffic accidents. The seemingly counterintuitive observation of higher accident rates during prevalent weather conditions like clear skies is noteworthy. The pronounced weekly seasonality in the accident data stands as a testament to the cyclic nature of these occurrences, necessitating heightened preventive measures during specific times. The modeling endeavors underscored an essential lesson: the choice of the model must resonate with the dataset's nuances. In this scenario, gradient boosting models like LightGBM outshone others, emphasizing the prudence of tailoring model selection to the dataset's inherent characteristics.

Harnessing the prowess of such models can be instrumental in devising proactive safety strategies, especially in the face of fluctuating weather conditions.

7.3 Recommendations

Based on the insights and conclusions derived from the project, the following recommendations are put forth for consideration in traffic management and accident prevention:

Weather-based Traffic Management: Given the established relationship between weather conditions and accident occurrences, traffic management authorities could employ adaptive traffic management strategies based on prevailing weather conditions, such as speed limit adjustments and traffic light timing modifications.

Predictive Interventions: The identified seasonal patterns and successful modeling efforts suggest the potential for developing predictive tools that can forecast accident risks based on weather forecasts and historical data, enabling proactive measures to mitigate accident risks.

Public Awareness: Raising public awareness regarding the impact of specific weather conditions on road safety can enhance driver caution and reduce accident risks. Communication of real-time accident risk assessments based on current weather conditions can be disseminated through traffic apps and navigation systems.

Infrastructure Improvements: The insights derived can inform infrastructural improvements and road maintenance activities, focusing on addressing the vulnerabilities revealed by specific weather conditions, enhancing road safety during adverse weather.

8 Limitations and Future Directions

8.1 Limitations

During my investigation into the relationship between climate conditions and traffic accidents, I recognize the following limitations:

- **Tourism Impact:** The impact of seasonal tourism influxes, which can lead to increased traffic and potentially more accidents, was not considered.
- **Road Maintenance:** Ongoing road constructions, repairs, and the general condition of roads, which can influence accident rates, were not a part of this study.

- **Traffic Congestion:** Factors such as traffic jams, rush hour patterns, and general traffic density, which are essential determinants of accident rates, were not explored.
- **Driving Behavior:** Local driving habits, adherence to traffic rules, or the prevalence of driving under influence can also influence accident rates but were beyond the scope of this study.
- **Vehicle Types:** The mix of vehicles on the roads, such as trucks versus cars or motorcycles, was not analyzed.
- **Public Transportation:** The role of public transportation efficiency and its utilization rate in influencing the number of vehicles and thus accidents on the road was not considered.

8.2 Future Directions

Given these limitations, future research could delve into these areas to provide a more comprehensive understanding of the determinants of traffic accidents. Integrating data on tourism, road conditions, and traffic congestion, among others, could offer more nuanced insights and refine predictive models. Further, exploring the impact of driving behaviors and the mix of vehicles on the roads might uncover more intricate patterns influencing accident rates.

8.3 Final Thoughts

The project underscores the crucial interplay between weather conditions and traffic accidents and illuminates the potential of predictive analytics in enhancing road safety. The recommendations provided, grounded in statistically validated insights, aim to offer practical solutions for reducing accident risks and fostering a safer and more responsive traffic environment. The practical implications of the project extend to diverse stakeholders, including traffic management authorities, policy-makers, and the general public, offering a pathway towards more informed and weather-resilient road safety strategies.

9 Improvements

Model Optimization: Further refinement and optimization of the model parameters can be performed to enhance the prediction accuracy. Advanced techniques like Bayesian optimization can be explored for hyperparameter tuning of machine learning models.

Feature Engineering: Additional features can be engineered from the available data, such as rolling averages and more complex lag features, to potentially improve model performance.

Data Enrichment: Incorporating more granular and diverse data, such as hourly weather conditions and real-time traffic data, can provide more nuanced insights and improve the predictive capabilities of the models.

Evaluation on Diverse Locations: Extending the analysis to include a wider array of cities with varying climatic conditions can offer more generalized and robust conclusions regarding the impact of weather on accidents.

10 Extensions

Inclusion of Additional Variables: Future analyses could explore the inclusion of additional weather-related variables like wind speed, humidity, and road condition to provide a more comprehensive view of the factors affecting accident incidence.

Application of Deep Learning Models: More advanced deep learning models, such as Transformers, can be experimented with for time-series forecasting to capture more complex patterns and relationships in the data.

Real-Time Prediction System: Developing a real-time prediction system integrating live weather data can enable dynamic risk assessment of road accidents, aiding in immediate and proactive traffic management responses.

11 Final Note

This documentation represents a structured, detailed, and coherent overview of the project "Analyzing Climate's Role in Accidents in Nashville, U.S.A.". It encapsulates the methodologies, analyses, insights, conclusions, and recommendations derived from the project, providing a comprehensive view of the undertaken work and its implications. The outlined future work and possible extensions offer avenues for further exploration and enhancement in the domain of weather-based accident analysis and prediction.

12 Acknowledgments

I would like to extend my gratitude to my mentors and peers at WBS Coding School for their continuous support, valuable feedback, and knowledge that they shared throughout the duration of this Data Science Bootcamp.

I would also like to acknowledge the providers of the datasets used in this project. Their efforts in collecting, cleaning, and sharing the data have enabled the exploration of meaning-

ful insights and the development of predictive models to understand the impact of weather conditions on traffic accidents.

Furthermore, I am grateful to the various open-source communities and forums online. The shared resources, discussions, and solutions available have been invaluable in resolving queries and gaining deeper insights into the methodologies and technologies used in this project.

Lastly, heartfelt thanks to my family and friends for their unwavering support and encouragement throughout this learning journey. Their belief in my capabilities and constant motivation has been a source of strength and inspiration.

This journey has been incredibly rewarding, and I am excited about applying the acquired knowledge and skills to contribute to the field of data science and to solve complex problems in the future.

References

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. *A Countrywide Traffic Accident Dataset.* 2019. Licensed under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. *Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.* In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[3] *Meteostat API Collection Methodology.* The dataset is sourced from the Meteostat API, utilizing Python libraries to ensure the accuracy and comprehensiveness of the information. Licensed under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0). https://www.kaggle.com/datasets/guillemservera/global-daily-climate-data