Understanding Seasonality in the Automotive Industry (Price Analysis)

A Project by Team 3 (Nabeel, Saksham, Riddhi, Spriha)

Summary

This report presents an in-depth study conducted by the University of Illinois Chicago's students of IDS 160 class - Team #3 to understand the factors influencing used car prices, focusing on weather pattern variations. Utilizing historical data from various sources, the team developed a predictive model to estimate used car prices.

Introduction

The used car market in the United States represents a significant segment of the automotive industry, characterized by its dynamic nature and the complex interplay of various factors that influence car prices. Recognizing the importance of this market, our project is centered around gaining a deeper understanding of these influencing factors, particularly focusing on the impact of weather and other critical variables on car prices.

We have chosen the regions of Georgia and Florida as our primary data sources for this study. These areas are strategically selected due to their unique climatic conditions, which vary significantly and thus provide a rich context for examining the influence of weather on car pricing. The variability in weather patterns, from intense rainfalls and storms to more temperate conditions, offers a diverse dataset for analysis. This geographical focus not only allows us to delve into regional market dynamics but also serves as an ideal backdrop for studying the broader implications of climatic factors in the used car market.

In our approach, we meticulously gather and analyze data, prioritizing the extraction of the most relevant features that could predict car prices. This includes a comprehensive examination of variables such as odometer readings, model types, and more. Beyond these traditional factors, we integrate detailed weather data into our analysis. This integration is pivotal to our study, as it allows us to explore how specific weather conditions – ranging from rainfall and storm occurrences to more subtle climatic variations – potentially impact the valuation of used cars.

By combining these diverse datasets, our project aims to create a holistic and nuanced understanding of the used car market. We seek to identify key patterns and correlations, which will not only enhance our knowledge of market dynamics but also provide valuable insights for stakeholders within the automotive industry. Through this endeavor, we aim to contribute to a more informed and strategic approach to pricing and marketing within the used car sector, particularly in the context of varying environmental and market conditions.

Objectives

In the automotive industry, the fluctuation of used car prices is influenced by various factors. The objective of this project is to analyze and understand the range of elements that affect the market value of pre-owned vehicles. This includes examining economic indicators and specific car features to determine how they impact pricing. The current market analysis often overlooks the complexity of these factors, and this project aims to fill that knowledge gap with thorough research.

Central to this project is the creation of a predictive model designed to estimate the prices of used cars. This model aims to incorporate the findings from our research, converting complex market data into a functional forecasting tool. The intention is to provide industry players with a resource for making informed decisions, supported by predictions based on rigorous data analysis that captures the intricacies of market behavior.

This objective will be pursued through a methodical process that leverages a variety of data, including historical sales data and climatic conditions, among others. The methodology combines statistical analysis with advanced machine learning techniques to create a model that accurately captures the patterns of price variation due to seasonality. The ultimate goal is to offer a detailed understanding of the pricing mechanisms in the used car market through this model.

Scope of Work

The scope of work for our project on understanding price seasonality in the automotive industry encompasses four primary areas: data collection, analysis, model development, and validation and testing.

Data Collection: Our first step is the aggregation of historical data related to used car pricing, which is fundamental to our analysis. This data collection phase extends to encompass weather conditions, supply chain metrics, and relevant market trends. By compiling a comprehensive dataset, we aim to create a robust foundation upon which our analysis will be built. The data collected will be scrutinized for quality, relevance, and integrity to ensure that subsequent phases of the project are based on accurate and reliable information.

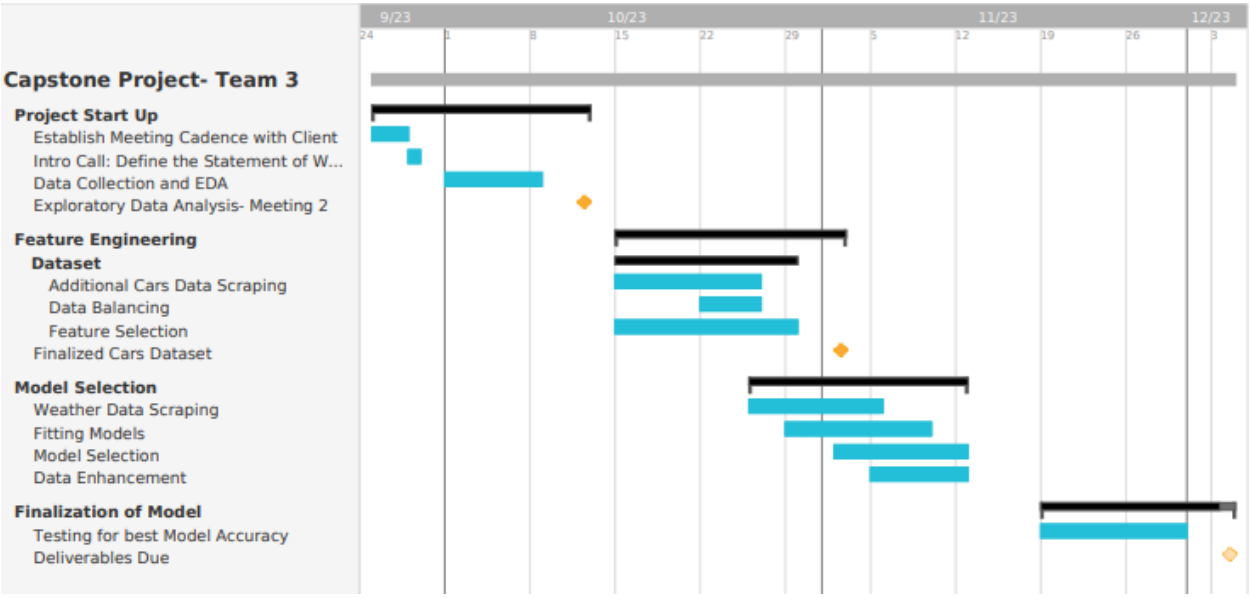
Analysis: With a rich dataset in hand, the next phase involves employing statistical analysis to decipher the significance and weight of each factor influencing car prices. This analytical phase is crucial as it allows us to understand not only the standalone impact of individual variables but also the interrelations between them. Through this meticulous analysis, we aim to identify patterns and trends that are pivotal in the pricing of used cars, thereby uncovering the nuanced interplay of the collected factors.

Model Development: Upon completing the analysis, we proceed to the development of a predictive model. This model is designed to use the insights gleaned from our analysis to forecast used car prices with high accuracy. The development process involves selecting appropriate algorithms and computational techniques that can best represent the complexities of the market data. Considerable effort will be invested in ensuring that the model is not only theoretically sound but also practical for real-world application.

Validation and Testing: The final phase of our scope involves rigorous validation and testing of the predictive model. This is a critical step where we deploy various testing and validation methods to verify the model's accuracy and reliability. The model will be subjected to a series of tests using new data sets to ensure it performs well across different scenarios and maintains a high level of prediction accuracy. This phase is iterative, with feedback loops enabling continuous refinement of the model for optimal performance.

Through these structured phases, we aim to deliver a predictive model that is a valuable asset for stakeholders in the automotive industry, facilitating better-informed decision-making regarding used car pricing.

Project Timelines



The Project Roadmap, as detailed in the Gantt chart, outlines the structured timeline and sequential phases planned for the successful completion of the project on price seasonality in the automotive industry. This schedule ensures that all team members and stakeholders are aware of the project's progression and key milestones.

The project commenced with a "Review of White Paper" from September 20 to October 1, where existing literature and preliminary data are examined to establish a foundational understanding of the subject matter. Following this, "Initial Meeting and Planning" took place between September 28 and October 4, setting the stage for project execution with clear objectives and roles.

Subsequently, "Data Collection and EDA" (Exploratory Data Analysis) was performed from October 5 to October 16. This phase involved gathering relevant data and performing initial analysis to identify patterns and outliers. "Feature Engineering", occurred from October 15 to October 25, a critical step where data was transformed and optimized for modeling.

"Advanced Data Analysis and Method Selection" was performed from October 25 to November 5. This stage involved applying sophisticated analytical methods to extract deeper insights from the data and selecting appropriate modeling techniques. Following this, "Data Enhancement" was performed from November 1 to November 15 with focus on improving the quality and richness of the data set.

The penultimate phase, "Finalizing the ML Model", was performed from November 16 to November 30. During this period, the machine learning models were refined, tested, and prepared for final deployment. Finally, "Project Delivery and Stakeholder Presentation" occurred from December 1 to December 5, where the project's outcomes and deliverables will be formally presented to the stakeholders.

Deliverables

The deliverables for this project have been clearly defined to encapsulate the breadth and depth of the research conducted on the seasonality of used car prices. These are itemized as follows:

1. ***Comprehensive Report***: The cornerstone deliverable is a detailed report that elucidates the factors influencing used car prices. It presents a thorough analysis of market dynamics and explicates the statistical model that has been developed and refined to predict car prices with high precision.

2. ***Predictive Model***: A critical deliverable is the predictive model itself, designed with the capability to forecast used car prices accurately. This model stands as a testament to the project's analytical rigor and practical application, providing stakeholders with a robust tool for price estimation.

3. ***Presentation***: The final deliverable is a comprehensive presentation that succinctly highlights the key findings of the study and outlines the recommendations derived from the analysis. This presentation is tailored to communicate the strategic insights to stakeholders effectively, ensuring that the implications of the research are clearly understood and actionable.

Data Sources

The project's data sources are as follows:

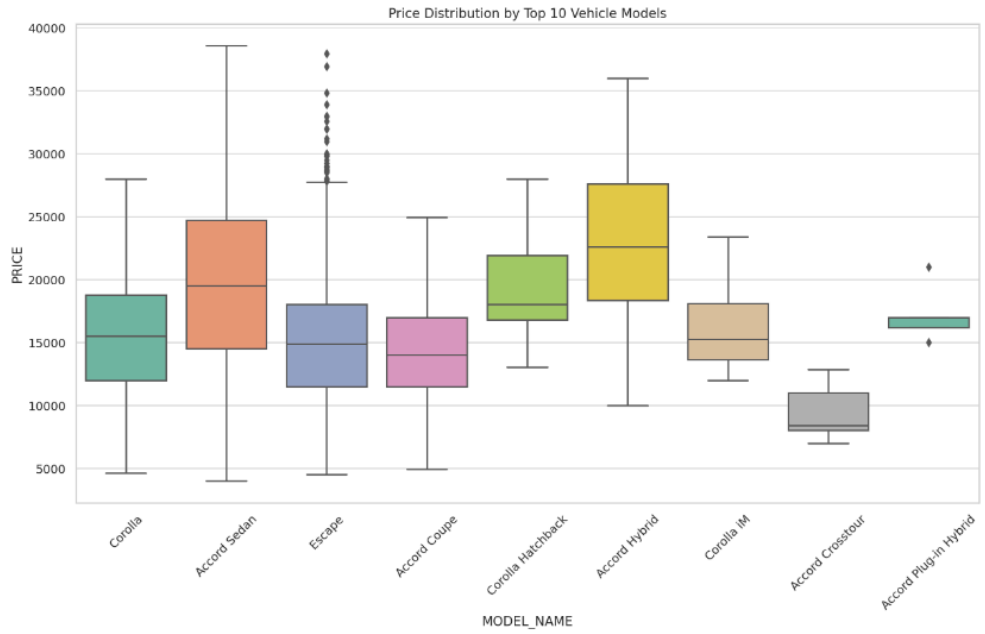
- ***National Centre for Environmental Information***: Provides historical climate data crucial for assessing weather impact on car prices.
- ***National Oceanic and Atmospheric Administration***: Offers atmospheric data that aids in correlating broader weather patterns with pricing trends.
- ***Cars.com***: Supplies real-time market data on car listings and prices, reflecting consumer demand and pricing fluctuations.
- ***2021 Dataset from CCC Intelligent Solutions***: Delivers specific automotive industry data, enriching the analysis with sector-focused insights.

The data has been processed and balanced using techniques from the Imbalanced-Learn (Imblearn) library to ensure that the samples are representative and to address any potential biases in the dataset. This methodological approach enhances the validity of the subsequent analysis and model development.

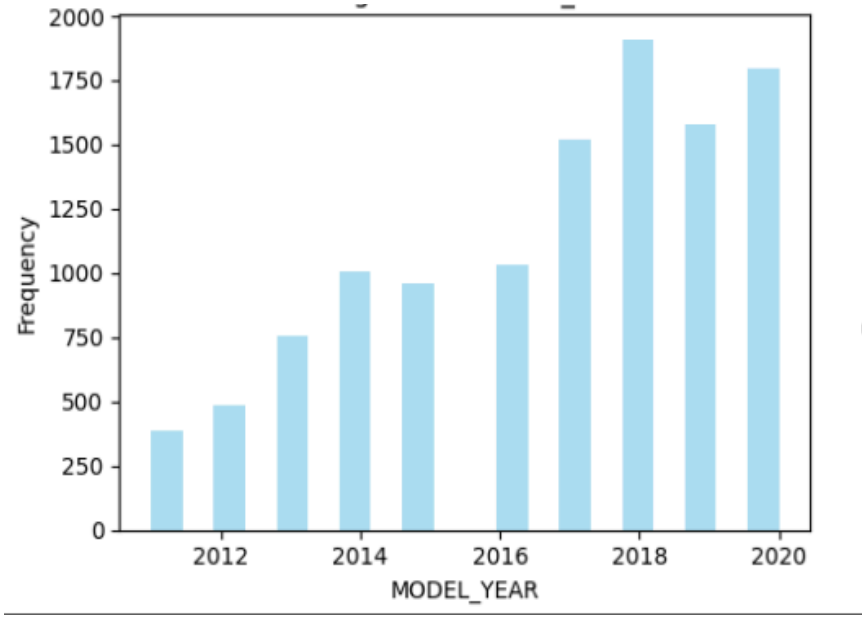
Exploratory Data Analysis

Price Distribution:

Box Plot

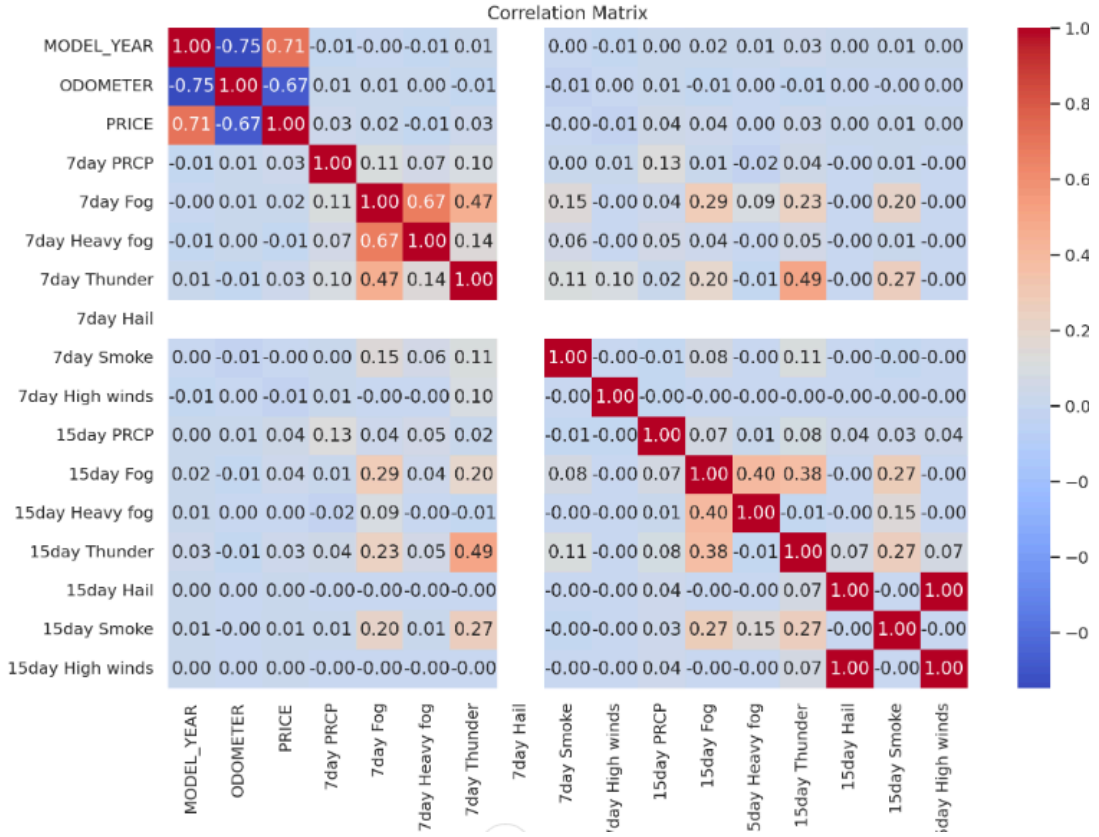


Bar Chart

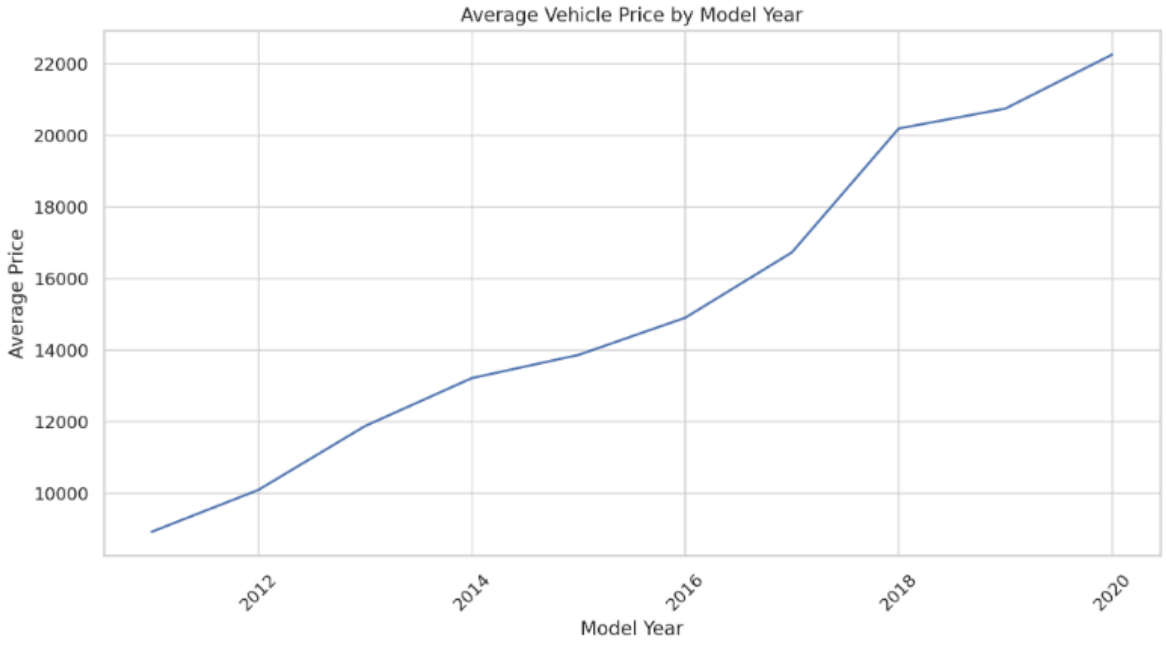


Scatter Plot

Correlation:



Price by Model Year:

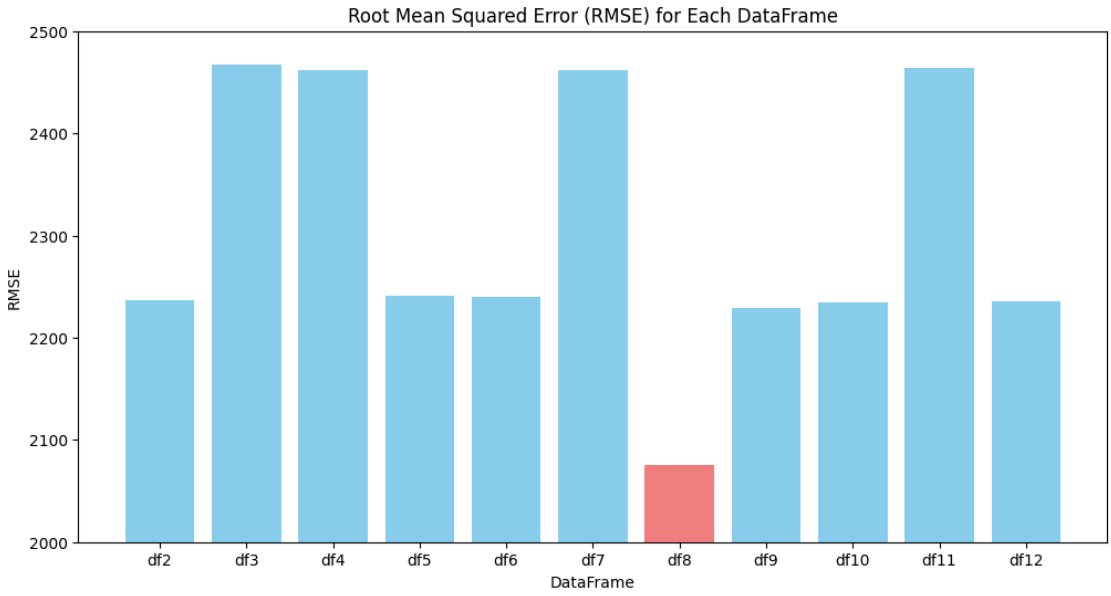


Adjusting Weather Dates:

This project sought to delve into the often-overlooked relationship between weather conditions and car prices, employing a novel approach by considering weather data from various time frames leading up to the sale date. Specifically, we investigated the effects of weather 7 days prior, 15 days prior, and 30 days prior to the actual sale date on the pricing dynamics of cars.
To facilitate a granular investigation, we segmented the data into three distinct time intervals: 7 days prior, 15 days prior, and 30 days prior to the sale date. Merging these datasets allowed us to draw correlations between weather conditions and car prices at different points in time.

Feature selection played a crucial role in refining the dataset. We identified key weather variables that were likely to influence car prices and meticulously handled missing or irrelevant data to maintain the integrity of the dataset. The subsequent implementation of regression models enabled us to predict car prices based on historical weather variables, with the evaluation of model performance relying on the Root Mean Squared Error (RMSE) metric.

ID_NUMB	MODEL_YI	PRICE	CITY	AD_DATE	0day PRCP	0day Fog	0day Heav	0day Thun	7day PRCP	7day Fog	7day Heav	7day Thun	15day PRC	15day F	15day Hea	15day Thu	30day PRC	30day Fog	30day Hea	30day Thu	30da
67	2014	14500	POMPANC	7/10/2021	0.49	1	0	1	0	0	0	0	0.18	1	0	1	0	0	0	0	0
70	2018	19500	ATLANTA	8/14/2021	0.43	1	0	1	0.41	1	0	1	0	1	0	1	0	1	0	1	0
193	2016	17777	ATLANTA	6/21/2021	0.04	0	0	0	0	0	0	1	0.29	1	0	1	0	1	0	0	0
249	2012	13659	SAVANNAI	10/4/2021	0	1	0	0	0	0	0	0	0.06	1	0	0	0	0	0	0	0
254	2017	12731	FORT MYE	1/27/2021	0.01	1	1	0	0	0	0	0	0.07	1	0	0	0	0	0	0	0
259	2019	19098	GAINESVIL	6/12/2021	0	0	0	1	0.09	0	0	0	1.01	1	0	1	0	0	0	0	0
285	2011	10491	ATLANTA	9/20/2021	0.51	1	0	0	0	1	0	0	0	1	0	0	0.06	1	0	0	0
308	2013	17300	HOLLYWO	9/30/2021	0.14	1	0	0	1.79	1	0	1	0.67	1	0	1	0	0	0	1	0
415	2019	30495	HOLLYWO	11/11/2021		0	0	0		0	0	0		1	0	0	0.08	0	0	0	0
437	2011	7980	MCDONOI	10/6/2021	1.45	0	0	0	0	0	0	0	0.3	1	0	0	0	0	0	0	0
453	2020	19990	ATLANTA	8/14/2021	0.43	1	0	1	0.41	1	0	1	0	1	0	1	0	1	0	1	0
479	2019	22596	ATLANTA	1/8/2021	0.16	1	0	0	1.02	1	1	0	1.25	1	0	0	0	0	0	0	0
488	2012	9924	CONYERS	1/13/2021		0	0	0		0	0	0		1	0	0	0.43	0	0	0	0
506	2020	31491	VALDOSTA	7/18/2021		0	0	0		0	0	0		1	0	0	0	0	0	0	0



df8 depicts a dataframe which has all the cars data and the corresponding weather data for 7 days and 15 days prior to the date of sale.

Model Fitting:

We trained the models initially on the training data and evaluated their performance on the test data. We'll use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as our evaluation metrics.

Model	Mean Absolute Error	Mean Squared Error	R-squared	RMSE
Linear Regression	1329.676	2969397.247	0.906748	1723.194
Random Forest	1029.477	1850617.560	0.941882	1360.374
Random Forest with Gridsearch	1038.137	1860517.083	0.941572	1364.008
XGBoost with Gridsearch	928.164	1497875.067	0.952960	1223.877
Neural Networks	N/A	N/A	N/A	2319.049

Interpretation

- XGBoost with Gridsearch has the lowest MAE and RMSE among the three models, making it the best performer in this comparison.
- Random forest also shows good performance, slightly better than Random Forest with Grid Search in terms of RMSE.
- Linear Regression, while simpler, has higher error metrics, indicating it might not capture the complexity of the data as well as the other models.
- Neural Networks also show a significantly high RMSE value suggesting further need of parameter tuning.

Next Steps

Having established a compelling connection between historical weather data and car prices, the next logical step in this research journey is to harness the predictive power of this relationship. This phase involves the development of a sophisticated tool that integrates weather forecasts to predict car prices accurately. By bridging the gap between meteorological predictions and pricing dynamics, this tool aims to empower stakeholders in the automotive industry with actionable insights for strategic decision-making.

The predictive tool will be implemented as a web-based application or software suite, accessible to users within the automotive industry. A secure and scalable cloud infrastructure will support the continuous integration of real-time weather forecasts, ensuring the tool's reliability and accuracy.

Validation of the predictive tool will involve comparing its predictions with actual car prices over time. Continuous monitoring and feedback from users will drive iterative improvements, allowing the tool to adapt to changing market dynamics and evolving weather patterns.

Having successfully developed a predictive tool that capitalizes on the correlation between weather conditions and car prices, the natural progression is to expand the scope from a regional level to a national scale. This initiative seeks to apply the insights gained from the Florida market to the entire United States, recognizing the diverse climates and market dynamics across the country. By broadening the geographical reach, the predictive tool can offer comprehensive and nuanced insights into how weather influences car prices on a national scale.

Challenges

The project encountered several challenges which are detailed below:

- Data Complexity: The multifaceted nature of the data, with its various dimensions and granularity, posed challenges in analysis and interpretation.

- Imbalanced Data: Disproportionate representation of data classes required advanced techniques to balance the datasets for accurate modeling.

- Limited Access to Resources: Constraints in accessing comprehensive datasets or advanced analytical tools impeded some aspects of the research.

- Feedback and Iteration Cycles: The need for continuous refinement of models based on feedback required substantial time and effort for iterative development.

- Technical Problem-Solving: Addressing and resolving complex technical issues was an ongoing challenge throughout the project lifecycle.