19.01.2023

CS210-Project

Final Report

Arda Hacıoğlu-31233

**Motivation**

Beginning this course, our professor underlined the importance of pursuing original and personal projects while avoiding the ordinary and the overdone. According to this advice, I set out a project to discover that would combine my academic objectives with my passions and interests outside of university. A combination of academic curiosity and personal connection led me to choose to use the data from my own diecast car collection for this project. Although owning an automobile in real life has obvious difficulties, diecast model collecting is not only possible but also highly enjoyable. This collection is not just a hobby for me, it is a carefully chosen repository of little marvels of automobile design, each with a unique backstory and importance. Using this collection for my project offers a special chance which gives me a new perspective in a field that is frequently dominated by more traditional datasets by allowing me to dive into a dataset that is both personally relevant and highly familiar. This strategy aims to present readers and project instructors with a unique and captivating perspective and through my analysis of this collection, I hope to reveal hidden patterns and ideas that offer a fresh perspective on the complexities of diecast model collecting experience.

**Data Source**

The source of the dataset for this project is my own diecast vehicle collection, which is a carefully selected collection of 745 different models. Data collection for each of these diecast vehicles was a complex and informative process.At first, the work was interpreting the data included on every model. These codes, which tend to be found on the underside of diecast automobiles, enable access to information about each model, including the manufacturer and model of the vehicle, its manufacturing year, and particular features.

However, not every model in my collection came with an instructive code. A significant portion of my collection consists of vintage models, which does not have an instructive code .This required me to conduct a thorough internet search. Every vehicle without a code posed a different set of difficulties, transforming the data collection procedure into a cross between historical research and detective work.Considering the vast volume of my collection, it was a challenging work to organize all of this information into a single, cohesive file. The project required careful cross-referencing and recordkeeping, which is indicative of the collection's richness and diversity. The process of assembling this dataset was quite satisfying, despite how difficult the task was. This was not just an exercise in data input; it was also a process of organizing a particular interest into a format that could be analyzed and analyzed again.

**Data Analysis**

The project's foundation is a dataset drawn from my personal diecast car collection, comprising diverse attributes such as manufacturers, colors, production years, and body types. This unique dataset helps to show the scope and development of the collection. During the essential preprocessing stage, managing missing values was the primary focus. Recognizing that these gaps might blur results, they were either removed or carefully restored using statistical measures such as means or medians, especially in important numerical aspects such as 'Production Year'. Converting categorical data into a numerical representation was another crucial step that was required for efficient model processing. One-hot encoding was used to do this, especially for features like "Car Brand" and "Body Type." We made sure that the diversity of the collection was appropriately represented by transforming categorical values into binary representations. This laid a strong basis for further exploratory data analysis, hypothesis testing, and machine learning modeling.

Exploratory data analysis (EDA) was a crucial stage in our effort to extract insights from the dataset of my diecast car collection. We used bar graphs to help us identify the most frequent brands and colors in the collection by examining the distribution of automobile brands (Manufacturer) and colors (Color). Histograms were crucial, particularly when analyzing the cars' production years (Production Year), which revealed popular time periods and production patterns. We also performed basic statistical studies, which yielded information on body types, commonly occurring car brands, and average manufacturing years. This offered a thorough understanding of the collection's general organization and evolutionary patterns. Finding significant trends and patterns was another aspect of the EDA process in addition to visualization, especially when it came to the popularity trajectories of various car models and types over time. We obtained further insights into the evolution of the collection by investigating the relationships across brands, models, and production years. This established a strong basis for further in-depth studies and hypothesis testing. This stage enabled us to comprehend and analyze the complex relationships included in our dataset, which was essential in determining the course of our research.

The main claim of our hypothesis is that the 'Coupe' body type remains popular irrespective of the manufacturer, color, and year of manufacturing. Based on interesting trends seen during the EDA phase, specifically the 'Coupe' body type's dominance across various colors and production years, this hypothesis was developed. The 'Coupe' models' continuous appeal was called into doubt by our exploratory study), which revealed an interesting pattern that led directly to the formation of this hypothesis. We used statistical techniques such as the T-examine and Chi-Square Test to examine this hypothesis.The relationship between the 'Coupe' body type and other factors like manufacturer, color, and production year has been determined in large part thanks to these tests. The interpretation of these test findings confirmed our idea and helped to explain the continued appeal of the 'Coupe' body type in my collection. Detailed hypothesis testing not only validated our first findings from the EDA, but it also gave our conclusions regarding the distinct position of the 'Coupe' body type in the dataset a more systematic and empirical foundation.

In our project, we used machine learning models—more especially, the Random Forest algorithm—to forecast the automobiles in the dataset's "Body Type." because Random Forest can handle complex data structures and has a high accuracy potential, it was selected. To forecast the "Body Type," the model was trained using features including "Manufacturer," "Color," and "Production Year." The dataset was divided into training and testing sets, and the model was trained on the former throughout the period of model training and optimization. To get the best performance, the model's hyperparameters were adjusted using techniques like Grid Search and Random Search. A variety of metrics, including accuracy, precision, recall, and F1-Score, were used to assess the model's performance on the test set. The important goal was to make reliably distinctions between “Coupe” and other body types.This assessment was essential in figuring out how well the machine learning model could distinguish between the 'Coupe' body type and other body types based on specific traits. The outcomes offered important information on how well the model predicted 'Coupe' body type, which strongly supported the validity of our theory. Our hypothesis regarding the specific qualities of the 'Coupe' body type within the collection was validated by the model's ability to reliably identify 'Coupe' body types based on distinguishing traits. This analysis and interpretation of the model results not only validated our hypothesis but also showcased the potential of machine learning in uncovering deeper patterns and insights within specialized datasets like our diecast car collection.

**Findings**

The journey through this project, particularly the exploration into my diecast car collection, has been revealing on multiple levels. A major accomplishment was confirming my hypothesis that the 'Coupe' body form remained popular despite changes in manufacturer, color, and manufacturing year. The unique distribution of the 'Coupe' model was highlighted by statistical tests such as the T-Test and Chi-Square Test, providing evidence in favor of this. The exploratory data analysis highlighted the 'Coupe's' unique location in my collection and provided more insight into its enduring appeal across eras and colors. My idea concerning Coupe's uniqueness was supported by the insights obtained from the Random Forest model, which skillfully separated it from other body types. These results went beyond simple statistics; they helped me comprehend the dynamics of my collection and the unique qualities of the 'Coupe' on a deeper level. This study was more than just a learning exercise; it was a journey into the heart of my interest, giving me an entirely new perspective on the subtleties present in my collection and a deep comprehension of it. Beyond the parameters of this work, the process taught me the importance of thorough study in discovering the hidden stories behind my interest.

Limitations and Future Work: This project, which is based on my own diecast car collection, provides interesting insights, it is not without limitations. Firstly, the data completeness and accuracy are limited. Secondly, the analysis in its current scope, primarily focuses on basic details like brand, model, and production year etc. It does not thoroughly explore nuanced details like historical significance, rarity, or market value of the models. Finally, the lack of detailed coding on some models necessitated reliance on online research, potentially introducing data inconsistencies and incompleteness. Furthermore, the representativeness of the dataset is restricted to the ones in my collection. The findings may not be fully generalizable as a result of this failure to fully reflect the global range of diecast automobile models.

Looking ahead to improve the vision of my project, a more representative and diversified sample might be obtained by adding more models to my collection. This extension would improve the volume of the dataset as well as its diversity with regard to the sorts of models and historical eras.

A larger and more comprehensive dataset would also be produced by adding further factors like the models' historical background, rarity, and current market value. Deeper insights could be obtained by using machine learning and advanced data analysis techniques, such as predictive modeling of market values or changes in the popularity of car models over time. Finally, this project presents an insightful look into the world of diecast car collecting, but it also creates many chances for growth and improvement. Future projects seek to expand the study's appeal to collectors, enthusiasts, and researchers in addition to improving its analytical depth.