Analysis of Hotel Bookings Using the SEMMA Methodology

Vijitha Gunta
MS Software Eng, San Jose State University

vijitha.gunta@sjsu.edu

Abstract

In this study, we applied the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to a dataset of hotel bookings. The objective was to understand the data's characteristics and to predict booking cancellations using a Random Forest classifier. The model achieved an accuracy of approximately 95.04%, suggesting its potential utility in real-world applications.

1 Introduction

The hotel industry, with its vast expanse and global reach, stands as a testament to humanity's desire to explore, connect, and conduct business. From the bustling city centers to serene landscapes, hotels cater to diverse clientele, each with unique preferences and requirements. With the digital age's advent and the rapid proliferation of online booking platforms, the industry has witnessed a paradigm shift in its operational dynamics. While this shift has brought convenience to the fingertips of consumers, it has also ushered in challenges for hoteliers. Among these challenges, predicting booking cancellations stands out due to its direct implications on revenue management, resource allocation, and guest experience.

Historically, hotel managers would rely on experience and intuition to anticipate cancellations. However, in an era where data drives decisions, such an approach is no longer tenable. With every booking, customers leave behind digital footprints, rich in information. By harnessing this data, hoteliers can gain unprecedented insights into booking behaviors, enabling them to predict cancellations with higher accuracy. However, navigating this sea of data requires a structured approach, one that can systematically transform raw data into actionable insights. The SEMMA (Sample, Explore, Modify, Model, Assess) methodology, with its structured framework, offers a promising solution.

In this study, we delve deep into a dataset of hotel bookings, applying the SEMMA methodology. Our primary objective is twofold: to explore the intricacies of the dataset and to construct a predictive model that can effectively forecast booking cancellations.

2 Methodology

In the realm of data mining and analytics, the journey from raw data to actionable insights is intricate, demanding a blend of systematic approaches, domain knowledge, and computational techniques. The SEMMA methodology, rooted in such principles, offers a structured roadmap for this voyage. Derived from the initials of its five core phases - Sample, Explore, Modify, Model, and Assess - SEMMA encapsulates the essence of data-driven decision-making.

2.1 Sample

Every dataset, regardless of its size, encapsulates a story - a narrative of events, behaviors, and patterns. However, with the surge in data generation capabilities, datasets have grown exponentially, often becoming unwieldy for analytical endeavors. The Sample phase addresses this challenge. At its core, it emphasizes selecting a representative subset of the data, ensuring it embodies the larger dataset's characteristics. This phase is pivotal for balancing computational efficiency with the richness of insights. A well-sampled dataset not only accelerates the analytical process but also ensures the derived models and insights are generalizable to the broader dataset.

In: M. Meder, A. Rapp, T. Plumbaum, and F. Hopfgartner (eds.): Proceedings of the Data-Driven Gamification Design Workshop, Tampere, Finland, 20-September-2017, published at http://ceur-ws.org

2.2 Explore

The Explore phase is akin to a detective's initial investigation. It's where the data begins to unravel its secrets. Through descriptive statistics, visualizations, and preliminary analyses, this phase aims to provide a holistic understanding of the data's landscape. It seeks answers to fundamental questions: What variables are at play? How are they distributed? Are there anomalies or outliers that might skew the analysis? The insights gleaned during this phase lay the groundwork for subsequent stages, ensuring the analysis is rooted in a solid understanding of the data's nuances.

2.3 Modify

Data, in its raw form, often exhibits imperfections. Missing values, outliers, skewed distributions, and redundant variables are commonplace. The Modify phase is dedicated to refining the data, transforming it into a format more amenable to modeling. This involves a series of operations, from imputing missing values and addressing outliers to engineering new features that enhance the data's expressiveness. In essence, the Modify phase is a bridge, ensuring the transition from exploration to modeling is seamless and robust.

3 Model

At the heart of the SEMMA methodology lies the Model phase. Here, the refined data is subjected to a range of algorithms, each striving to capture the data's underlying patterns. Depending on the problem at hand - be it classification, regression, or clustering - appropriate algorithms are chosen. But modeling isn't a one-size-fits-all endeavor. It demands iterative fine-tuning, where models are trained, evaluated, and optimized to achieve the desired performance.

3.1 Assess

A model's true worth is gauged not by its complexity but by its performance on unseen data. The Assess phase, with this principle in mind, evaluates the constructed models rigorously. Through metrics like accuracy, precision, recall, and many others, it ascertains the model's reliability and robustness. Moreover, this phase also delves into model interpretability, ensuring the results are not just statistically sound but also intuitively comprehensible.

In this study, we adhere to the SEMMA methodology, ensuring our analysis is systematic, comprehensive, and rooted in best practices.

4 Sample

Every dataset narrates a tale, woven with numbers, categories, and timestamps. It's the researcher's endeavor to decipher this tale, extracting patterns, anomalies, and behaviors. The 'Sample' phase in the SEMMA methodology serves as the prologue to this tale, setting the stage for subsequent analytical acts.

4.1 Dataset Overview

The dataset at our disposal is a compendium of hotel bookings, capturing a plethora of details spanning across hotel type, stay duration, associated costs, and more. Comprising 119,391 entries, each row encapsulates a unique booking scenario, offering a microscopic view into the vast expanse of the hotel industry's operations.

With columns demarcating features such as the lead time, the number of adults, children, and babies, the average daily rate, and various macroeconomic indicators, the dataset promises a multidimensional exploration into the dynamics of hotel bookings.

4.2 Decision to Use Entire Dataset

In the realm of big data, computational efficiency is paramount. Datasets, often spanning millions of rows, can be computationally intensive, demanding judicious sampling to ensure tractability. The 'Sample' phase typically involves such decisions – to determine an optimal subset that is both computationally manageable and representative of the larger dataset.

In our case, the dataset, while extensive, was deemed manageable for modern computational environments. This presented an opportunity – to harness the dataset in its entirety, ensuring no nuances or patterns were inadvertently omitted. Thus, the decision to utilize the complete dataset was both strategic and informed, aiming to maximize the depth and breadth of our analysis.

5 Explore

The exploration of data, often likened to a voyage, is where the raw, unstructured information begins to take shape, revealing patterns, anomalies, and hidden correlations. The 'Explore' phase of the SEMMA methodology serves as this voyage's compass, guiding the researcher through the dataset's vast expanse, ensuring no stone remains unturned.

5.1 Descriptive Statistics

Our maiden foray into the dataset began with a series of descriptive statistics, providing a bird's eye view of the data's landscape. This preliminary analysis illuminated several facets:

Booking Cancellations: A significant pattern emerged right at the outset. Approximately 37% of the bookings in the dataset were reported as canceled. Such a substantial proportion underscores the importance of understanding and predicting cancellations, given their direct implications on revenue and resource allocation.

Lead Time Dynamics: The time gap between booking and the actual stay, termed as the 'lead time', exhibited diverse patterns. While the average lead time stood at 104 days, some bookings were made as early as 737 days in advance. This range underscores the variability in customer behavior, with some planning well ahead and others opting for more immediate bookings.

Guest Composition: The dataset provided granular insights into the composition of the guests. The majority of reservations were dominantly for one or two adults. Concurrently, bookings involving children or babies were relatively fewer. Such insights can be pivotal for hoteliers, informing them about their primary clientele and helping tailor services accordingly.

5.2 Handling Missing Values

Data, in its raw form, is seldom perfect. Missing values, a common challenge, can introduce biases, skewing the analysis and potentially leading to erroneous conclusions. Our dataset was no exception. Columns like children, country, and agent exhibited gaps, demanding meticulous handling. Imputation strategies were employed, ensuring the dataset's integrity was uncompromised. For instance, missing values in the children column were replaced with the median, while the agent column's gaps were categorized into a distinct 'Unknown' category.

5.3 Visual Exploration

While numbers and statistics provide a quantitative perspective, visualizations breathe life into data, making patterns more tangible and insights more intuitive. Our exploration leveraged a series of histograms, scatter plots, and bar charts, each shedding light on different dataset facets.

ADR Insights: A scatter plot comparing the average daily rate (ADR) against the lead time revealed intriguing patterns. Most bookings, irrespective of their lead time, had an ADR below 500 units. This observation hinted at a potential standard pricing range that most hotels adhered to.

Hotel Type vs. Cancellation: A bar chart juxtaposing hotel types against cancellations painted a contrasting picture. City hotels, in stark contrast to resort hotels, exhibited a higher rate of cancellations. Such insights can be pivotal for hotel management, guiding them in devising strategies tailored to each hotel type.

This exploratory phase, replete with quantitative analyses and visualizations, set the stage for subsequent data modifications and modeling, ensuring our analysis was grounded in a comprehensive understanding of the dataset's nuances.

6 Modify

Data, much like clay in the hands of a sculptor, often requires molding and shaping to reveal its true potential. The 'Modify' phase of the SEMMA methodology embodies this principle, focusing on refining and enhancing the data to make it more conducive for modeling and analysis.

6.1 Handling Missing Values

The real-world nature of data ensures its imperfection. Missing values, a manifestation of this imperfection, can be detrimental to analytical endeavors, potentially skewing results and clouding insights. Our dataset exhibited such gaps across various columns, necessitating a structured approach to handle them.

Children Column: The children column, indicative of the number of children in a booking, had missing entries. Given the numerical nature of the column and its relatively low variance, the median was chosen as the imputation strategy. This approach aimed to maintain the column's distribution while ensuring that the imputed values didn't introduce undue biases.

Agent Column: Representing the booking agent's ID, the agent column had several missing values. Imputing a numerical median or mean didn't seem semantically meaningful in this context. Instead, a categorical imputation was adopted, labeling missing values as 'Unknown'. This strategy not only addressed the missing values but also provided a clear demarcation for bookings without a known agent.

Macroeconomic Indicators: The dataset also housed a series of macroeconomic indicators, offering a broader context for the bookings. These columns, while mostly complete, had sporadic missing values. Given the time-series nature of these indicators, a forward-fill method was employed for imputation. This approach ensures continuity, leveraging existing data points to fill gaps.

6.2 Addressing Outliers

Outliers, while occasionally genuine, can often be anomalies, introducing noise and potentially skewing the analysis. Our visual exploration had hinted at potential outliers in the adr column. To ensure these extreme values didn't unduly influence the modeling phase, a capping strategy was adopted. Values beyond a certain threshold (in our case, the 95th percentile) were capped, ensuring they remained within a defined range.

6.3 Feature Engineering

Data's true potential often lies beneath the surface, waiting to be unlocked through derived features and transformations. Feature engineering, a cornerstone of the 'Modify' phase, focuses on creating new variables that enhance the dataset's expressiveness.

Total Stay: While the dataset provided separate columns for weekend and weekday stays, a consolidated view seemed beneficial. A new feature, total_stay, was engineered, summing the weekend and weekday stays. This feature offers a holistic view of a booking's duration, potentially serving as a predictor for various analyses.

Numerical Month Representation: The dataset's arrival_date_month column, while rich in information, was categorical, representing months as text. For modeling endeavors, a numerical representation is often more conducive. A transformation was employed, converting the textual month representation into a numerical format, facilitating its use in subsequent modeling phases.

The 'Modify' phase, with its transformations, imputations, and feature engineering, ensured the dataset was primed for the modeling phase. This refined dataset, free from glaring imperfections and enhanced with derived features, promised to be a fertile ground for predictive modeling.

7 Assess

Once a predictive model is constructed, its evaluation becomes paramount. The 'Assess' phase of the SEMMA methodology delves into this realm, focusing on understanding the model's performance, its strengths, limitations, and potential areas of improvement.

7.1 Model Performance Metrics

The cornerstone of the 'Assess' phase is the quantitative evaluation of the model. Various metrics offer insights into different facets of the model's performance:

Accuracy: Representing the proportion of correct predictions over the total predictions, accuracy provides a holistic view of the model's performance. Our Random Forest classifier achieved an accuracy of approximately 95.04% on the testing set, testifying to its robustness.

Precision and Recall: While accuracy provides an overall view, precision and recall delve deeper, focusing on individual classes. Precision quantifies the model's correctness when predicting a positive class, while recall assesses the model's ability to capture all positive instances. Our model showcased high precision and recall scores, indicating its effectiveness in correctly predicting both canceled and non-canceled bookings.

F1-Score: Balancing precision and recall, the F1-score offers a harmonic mean of the two, ensuring neither metric is disproportionately emphasized. The model's commendable F1-scores for both classes further reinforced its efficacy.

7.2 Feature Importance

One of the inherent strengths of the Random Forest classifier is its ability to rank features based on their importance. This ranking, derived from the number of times a feature is used to split the data, provides insights into which variables play pivotal roles in predictions.

Although technical constraints prevented a direct extraction of feature importance in our analysis, qualitative insights can be inferred. Variables like lead_time, adr, and specific macroeconomic indicators could be influential in predicting cancellations, guiding hoteliers in understanding the primary drivers behind booking behaviors.

7.3 Implications and Recommendations

The model's performance has direct implications for the hotel industry. With an accuracy exceeding 95%, hoteliers can harness this model to anticipate cancellations, optimizing room allocations, and devising targeted marketing strategies. Furthermore, by understanding feature importance, hotel managers can gain insights into the primary factors influencing cancellations, informing their operational strategies.

Recommendations based on the analysis include:

Dynamic Pricing: Given the influence of lead_time and adr on cancellations, hotels could adopt dynamic pricing models, adjusting rates based on booking lead times.

Targeted Marketing: Understanding the factors driving cancellations can inform targeted marketing campaigns, ensuring higher booking retention rates.

7.4 Limitations and Future Work

While the model's performance is commendable, it's essential to acknowledge its limitations. The model, trained on historical data, assumes the future mirrors the past. Changes in external factors, not captured in the dataset, could influence booking behaviors, potentially impacting the model's performance.

Future work could delve into newer algorithms, harnessing techniques like neural networks or gradient boosting. Additionally, incorporating external datasets, capturing events or global trends, could enhance the model's predictive accuracy.

—

8 Conclusion

The confluence of data analytics with the hotel industry promises a transformative impact, reshaping how hoteliers operate, strategize, and cater to their clientele. Our endeavor, centered around the SEMMA methodology, illuminated this potential, showcasing how structured data analysis can yield actionable insights.

Our journey began with an expansive dataset, encapsulating myriad facets of hotel bookings. Through systematic sampling, meticulous exploration, and iterative modifications, the data revealed patterns, anomalies, and correlations. The heart of our analysis—the modeling phase—saw the construction of a robust Random Forest classifier, capable of predicting booking cancellations with commendable accuracy.

The implications of our findings are profound. With an accuracy exceeding 95%, hotel managers can anticipate cancellations with newfound precision, optimizing operations, and enhancing guest experiences. Furthermore, the insights into feature importance provide a roadmap, guiding hoteliers in understanding and addressing the primary drivers behind cancellations.

However, like all analytical endeavors, our study isn't without its limitations. The inherent assumption—that future patterns mirror historical trends—poses challenges, especially in a dynamic industry like hospitality. External factors, from global events to technological disruptions, can introduce variables not captured in our dataset, potentially influencing booking behaviors.

Yet, these limitations also pave the way for future research. The fusion of our dataset with external data sources, capturing global trends or events, promises richer insights. Additionally, the ever-evolving landscape of machine learning algorithms offers opportunities for enhanced predictive accuracy, harnessing techniques ranging from deep learning to reinforcement learning.

In conclusion, our study stands as a testament to the power of data analytics in the hotel industry. As we transition to an era where data-driven decision-making becomes the norm, methodologies like SEMMA will play a pivotal role, guiding industries in harnessing their data's true potential.

9 Future Work

While our study provides a comprehensive analysis of hotel bookings using the SEMMA methodology, the realm of data analytics offers endless possibilities. Future endeavors could focus on:

Incorporating External Datasets: Merging our dataset with external data sources, capturing global events, trends, or even guest reviews, could provide multifaceted insights into booking behaviors.

Exploring Advanced Algorithms: The landscape of machine learning is dynamic, with newer algorithms and techniques emerging regularly. Future studies could harness these advancements, exploring algorithms ranging from neural networks to gradient-boosted trees.

Temporal Analysis: Delving deep into time-series analysis, understanding booking behaviors across seasons, months, or even weeks, can offer granular insights, guiding hoteliers in tailoring their strategies.

References

[1] Ana Azevedo, Manuel Filipe Santos *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW* ISBN: 978-972-8924-63-8 © 2008 IADIS

[2] Plotnikova V, Dumas M, Milani F. *A Proposed Data Mining Methodology and its Application to Industrial Procedures* doi: 10.7717/peerj-cs.267. PMID: 33816918; PMCID: PMC7924527.