AIRBNB ANALYSIS

1. INTRODUCTION

In data-driven decision-making, Airbnb, a prominent player in the global hospitality industry, faces the ongoing challenge of optimizing user experience. A key aspect of this challenge revolves around efficiently guiding new users to suitable accommodations, thereby enhancing user satisfaction and engagement. Central to addressing this challenge is the ability to accurately predict a new user's first booking destination based on their initial interactions with the platform.

Extensive research within the realm of predictive analytics has underscored the importance of accurately forecasting user behavior for enhancing user experience and business operations. Studies such as those by Le and Naveed (2020) and Blecharczyk, Gebbia, and Chesky (2017) have demonstrated the potential impact of predictive modeling on user engagement and conversion rates within the hospitality industry. By leveraging advanced machine learning techniques, these studies have shown that personalized recommendations and targeted marketing strategies can significantly improve the likelihood of user conversion.

Against this backdrop, Airbnb has embarked on a Kaggle challenge, presenting a compelling question to the data science community: Can we predict a new user's first booking destination based on their initial interactions with the platform? This challenge not only reflects Airbnb's commitment to innovation but also aligns with its broader mission to enhance user experience through personalization and optimized marketing strategies.

The primary objective of this project is to address this challenge by harnessing the power of machine learning and advanced analytics. By meticulously analyzing user demographics, web session records, and other relevant data, we aim to develop predictive models capable of accurately forecasting new users' booking destinations. Such models hold immense potential for revolutionizing how Airbnb interacts with its users, enabling personalized marketing campaigns, optimizing resource allocation, and ultimately reducing the time between a user's initial interaction and their first booking.

This introduction sets the stage for our investigation, highlighting the significance of the problem at hand and providing evidence of a literature review that has informed our approach. Through rigorous analysis and modeling, we seek to not only enhance Airbnb's operational efficiency but also elevate the overall user experience, positioning Airbnb at the forefront of the hospitality industry's digital transformation.

2. METHODOLOGY

2.1 Dataset Background:

The dataset utilized in this project is sourced from Airbnb through a Kaggle competition, tailored to aid in predicting a new user's first booking destination by analyzing various user-related attributes. With 213,452 rows and 15 columns, each row represents a unique user, and each column denotes a specific attribute of these users. This dataset encompasses a broad spectrum of information ranging from basic user demographics to their interactions with the Airbnb platform.

Key attributes included in the dataset are:

- **ID**: A unique identifier for each user.
- **Date Account Created:** The date when the user's account was created on Airbnb.
- **Timestamp First Active**: A Unix timestamp representing the user's first interaction with the Airbnb platform.
- **Date First Booking**: The date when the user made their first booking, crucial for identifying the conversion rate and timing.
- **Gender**: The reported gender of the user, influencing the type of accommodations preferred.
- **Age**: User’s age, offering insights into demographic segments more active or inclined towards certain destinations.
- **Signup Method**: How the user signed up (e.g., through Facebook, Google, or directly on the Airbnb website), impacting user engagement.
- **Signup Flow**: The page from which a user came to the signup page, indicative of the effectiveness of different marketing strategies.
- **Language**: Default language set by the user, important for personalizing user experience and marketing communications.
- **Affiliate Channel**: Type of paid marketing bringing the user to the website, like direct, seo, or other.
- **Affiliate Provider**: The affiliate network or partners bringing the user, such as Google, Craigslist, or direct.
- **First Affiliate Tracked**: The first marketing effort the user interacted with before signing up, aiding in understanding effective touchpoints.
- **Signup App**: The application (Web, iOS, Android) used by the user to sign up, helpful for tailoring app-specific features or advertisements.
- **First Device Type**: Device type on which the user signed up, like Mac Desktop, iPhone, or Android Phone.
- **First Browser**: Browser used by the user during their first interaction with the platform.

- **Country Destination**: The country where the user made their first booking, categorized into several classes, including 'US', 'FR', 'CA', and 'NDF' (no destination found), serving as the target variable.

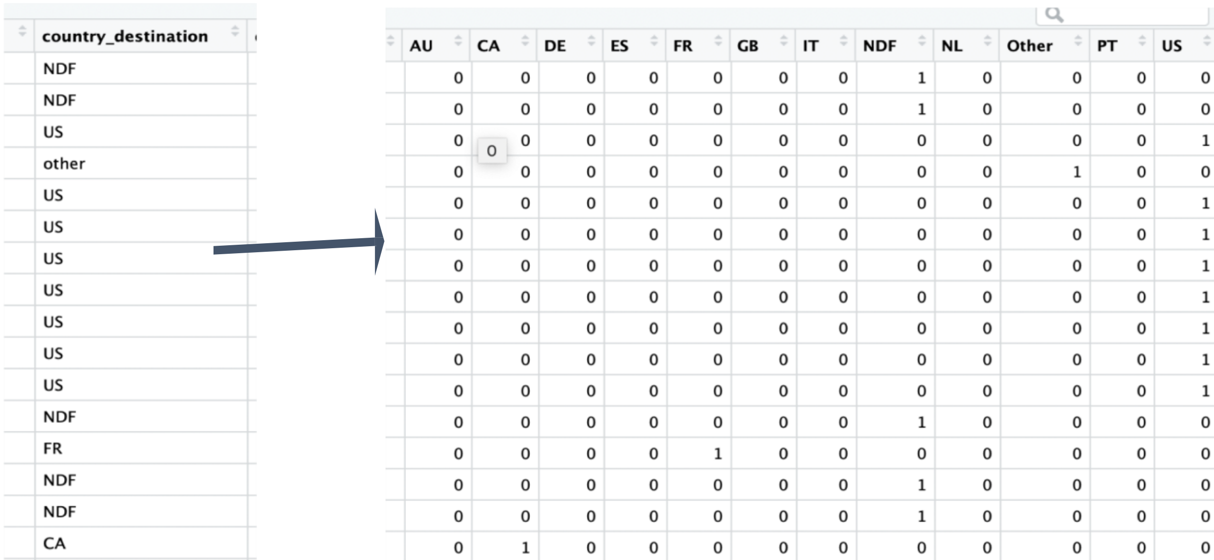
This dataset's significance lies in its diverse range of data points, providing a robust foundation for deep analysis aimed at accurately predicting first booking destinations. Understanding user demographics, behaviors, and platform interactions is critical for customizing user experiences and optimizing Airbnb's marketing strategies. Despite being relatively clean, the dataset may contain challenges such as missing values, particularly in the 'Date First Booking', 'Age', and 'Gender' columns, which need to be addressed through appropriate data preprocessing methods to ensure the reliability and robustness of subsequent analyses. Overall, this dataset not only enables a profound comprehension of user behaviors and preferences but also challenges analysts to refine their analytical techniques for effective outcome prediction, thereby aiding Airbnb in enhancing user engagement strategies and overall service delivery.

2.2 Data Preprocessing and Data Cleaning:

The process of preparing the dataset for analysis and subsequent one-hot encoding involves several steps to ensure data cleanliness and suitability for modeling. Initially, we have loaded the R packages such as tidyverse *Wickham et al., (2019).*, ggplot2 *Wickham, H. (2016),* data.table *Dowle, M., & Srinivasan, A. (2021),* random forest *Liaw, A., & Wiener, M. (2002).* The dataset is loaded using the `read.csv` function, facilitating its conversion into a dataframe named `df`. To streamline the dataset, unnecessary columns such as `date_first_booking` are removed, focusing only on relevant data. Missing values within the dataset are handled by replacing them with the mean of their respective columns, ensuring data integrity without introducing bias. Additionally, the `age` column is converted to an integer data type to facilitate analysis.

To better understand temporal trends, the `date_account_created` column is dissected into its year, month, and day components, stored in newly created columns (`dac_year`, `dac_month`, `dac_day`). Similarly, if the `timestamp_first_active` column exists, it undergoes the same processing to extract temporal information. In the absence of this column, a message indicating its non-existence is printed for clarity.

Utilizing one-hot encoding, the categorical variable `country_destination` is converted into binary variables representing each country destination, enhancing the dataset's compatibility with machine learning algorithms. Post one-hot encoding, the generated columns are renamed to reflect the respective country codes clearly, aiding interpretability during analysis.



country_destination
NDF
NDF
US
other
US
US
US
US
US
US
US
NDF
FR
NDF
NDF
CA

AU	CA	DE	ES	FR	GB	IT	NDF	NL	Other	PT	US
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0

In culmination, a final dataframe named `df_final` is constructed by excluding the original categorical column, now represented by one-hot encoding. This meticulous data preprocessing lays the groundwork for subsequent analysis and modeling tasks.

2.3. Machine Learning Models:

2.3.1. Selection of Machine Learning Models:

In our endeavor to predict the first booking destination of new users on Airbnb, we embarked on a comprehensive exploration of machine learning models. Our approach was guided by a meticulous selection process, considering the suitability of various models for the classification task posed by the Airbnb dataset. Drawing from the extensive literature on machine learning and classification algorithms, we identified Naive Bayes, K-Nearest Neighbors (KNN), and Random Forest as promising candidates, each offering distinct advantages in handling large datasets, processing multiple features, and providing valuable insights.

2.3.2. Implementation of Machine Learning Models:

Naive Bayes, renowned for its simplicity and efficiency, was chosen as a foundational model due to its ability to handle large datasets with multiple features. The decision to include Naive Bayes was informed by its widespread use as a baseline model in classification tasks, allowing

for a quick assessment of feature-target relationships. Our implementation of Naive Bayes involved utilizing the entire dataset for training, ensuring that the model captured all underlying patterns without succumbing to overfitting.

KNN, with its intuitive approach of making predictions based on the proximity of data points in the feature space, emerged as another compelling choice. This model was particularly suitable for our dataset, where users with similar characteristics often exhibit similar booking preferences. Leveraging insights from the literature on KNN, we tested the model using a subset of the data, focusing on key attributes such as gender, age, and user activity times to enhance its sensitivity to influential features.

The inclusion of Random Forest, known for its robustness and accuracy in classification tasks, further enriched our model ensemble. With its ensemble approach and feature ranking capabilities, Random Forest offered a promising avenue for handling the high-dimensional nature of our dataset while mitigating overfitting. We meticulously configured the Random Forest model, initially applying it to a subset of the data and subsequently scaling it to the entire dataset, with careful parameter tuning to strike a balance between bias and variance.

2.3.3. Evaluation of Machine Learning Models:

Our model evaluation process was anchored in established methodologies from the literature, encompassing cross-validation techniques, holdout sets, and performance metrics such as accuracy, precision, recall, and F1-score. Drawing inspiration from seminal works on model evaluation, we employed confusion matrices to visually assess each model's performance in correctly predicting booking destinations. Furthermore, feature importance analysis, particularly with Random Forest, provided valuable insights into the variables most influential in predicting user behavior.

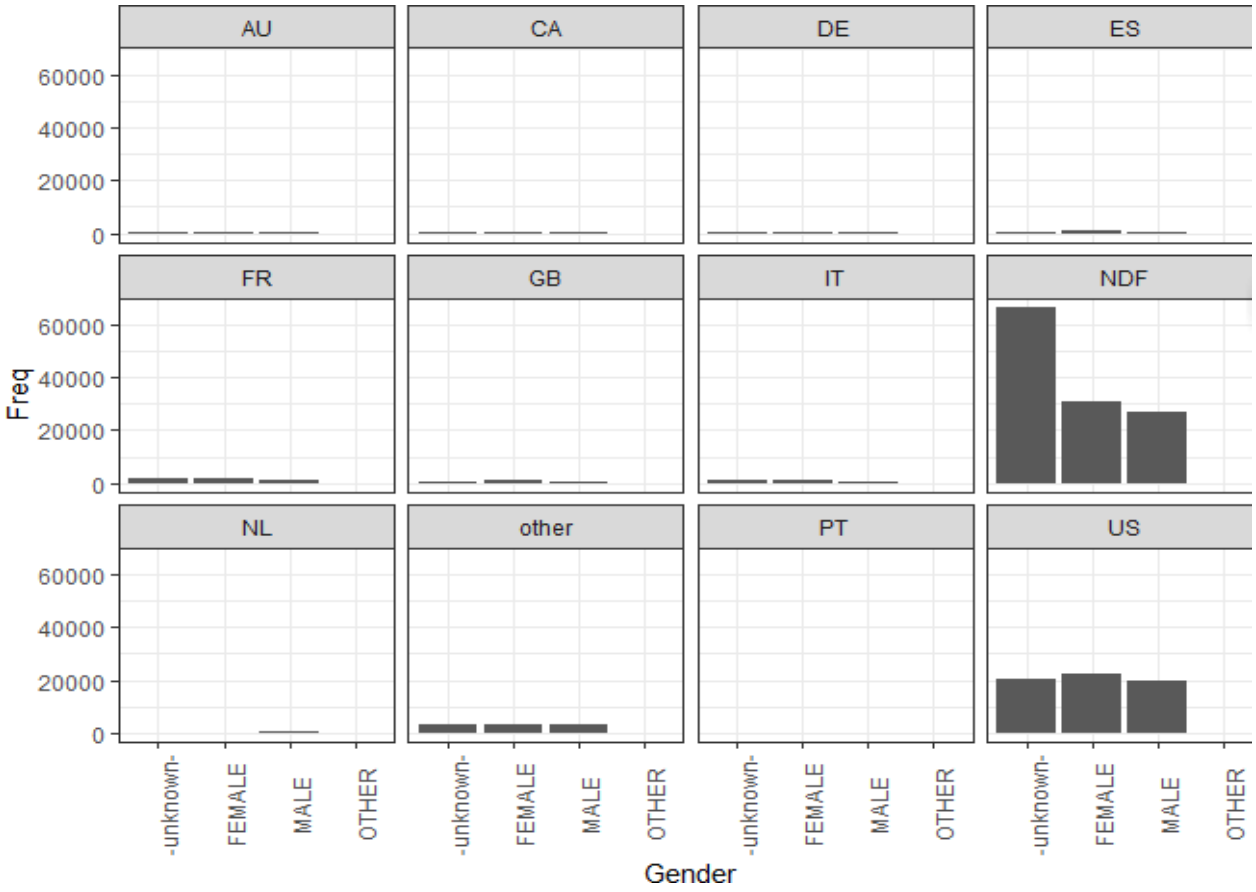
The findings and methodologies employed in our project are deeply rooted in the vast body of literature on machine learning and classification algorithms. References to seminal works by researchers such as Hastie, Tibshirani, and Friedman (2009) on Random Forest, and Manning, Raghavan, and Schütze (2008) on Naive Bayes, among others, informed our understanding and implementation of these models. Additionally, insights from Kaggle competitions and academic papers on Airbnb user behavior further enriched our approach, ensuring that our models were well-equipped to tackle the complexities of the dataset.

3. RESULTS

3.1. Exploratory Data Analysis: Unveiling User Behavior on Airbnb

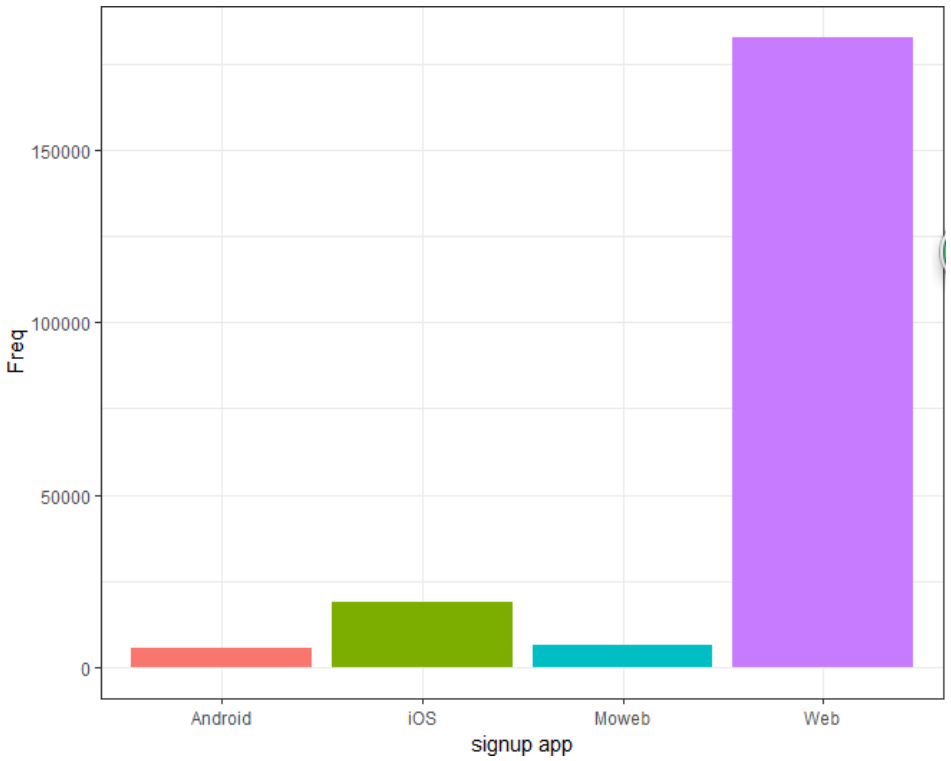
3.1.1. Gender Distribution Across Destinations:

Understanding the gender composition of Airbnb users and its variation by booking destination. The analysis unveiled a predominant booking trend among females, particularly in the United States. Additionally, a significant number of bookings were categorized as 'NDF' (no destination found), indicating instances where users did not complete a booking.



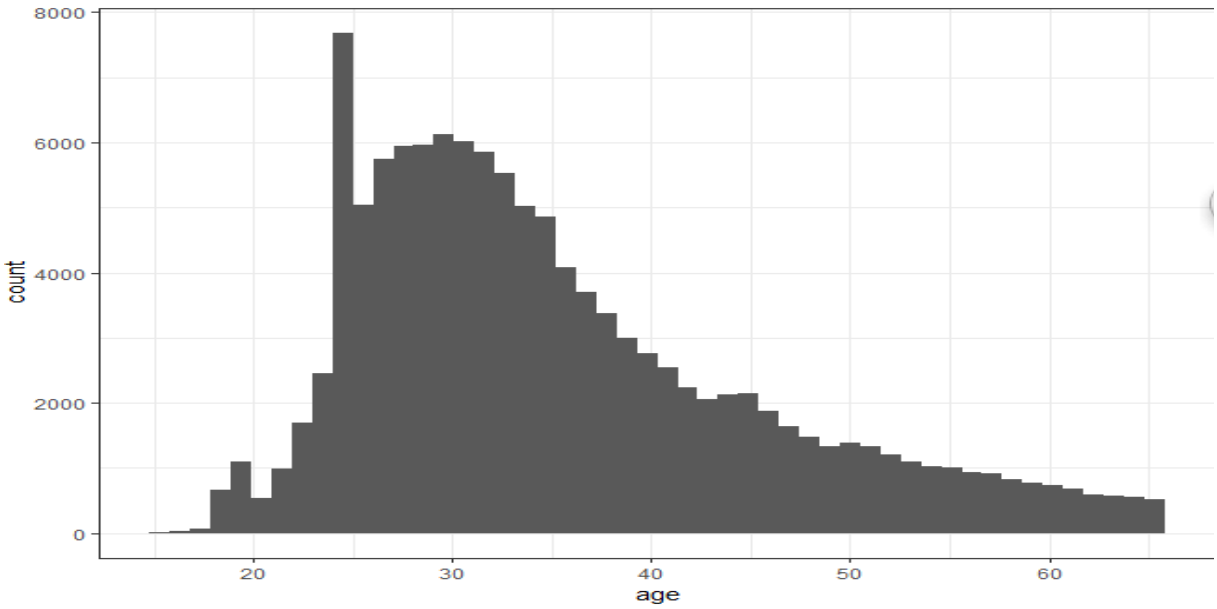
3.1.2. Signup Methods: Impact on Booking Preferences:

Identifying the most popular signup methods used by users to book accommodations on Airbnb. The majority of users prefer to book through the Airbnb website using web browsers like Safari and Chrome. This highlights the importance of a user-friendly and efficient online platform in facilitating bookings.



3.1.3. Age Distribution of Users: Tailoring Marketing Strategies:

Determining the age distribution of Airbnb users to tailor marketing and accommodation options effectively. The analysis revealed that a significant portion of Airbnb users falls within the 35-40 age range. This demographic insight can guide the customization of offerings to cater to specific age groups preferences.



3.2.Model Performance:

3.2.1. Naive Bayes:

Naive Bayes was chosen for its simplicity and efficiency in handling large datasets with multiple features. It serves as a baseline model, offering a quick assessment of relationships between features and the target variable. The Naive Bayes model was trained using the entire dataset to ensure capturing all underlying patterns without overfitting specific subsets of the data. The Naive Bayes model achieved an accuracy of 68.08%, serving as a benchmark for subsequent models.

```
      US
NBayes     0     1
     0 26514  9822
     1  3805  2550
> NB_wrong<-sum(category!=test$US )
> NB_error_rate<-NB_wrong/length(category)
> NB_error_rate
[1] 0.3192008
> accuracy <- (1-NB_error_rate)*100
> accuracy
[1] 68.07992
>
```

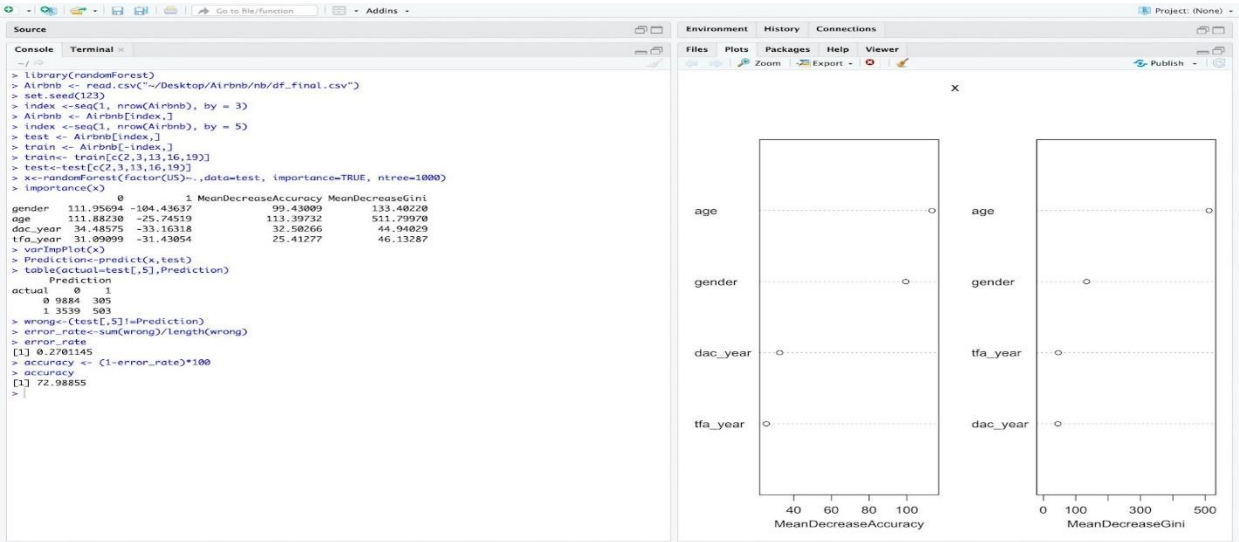
3.2.2. K-Nearest Neighbors (KNN):

KNN was selected due to its ability to make predictions based on the proximity of data points in the feature space, making it suitable for datasets where similar users often choose similar destinations. The KNN model was tested using 33% of the data, focusing on key attributes such as gender, age, and user activity times to refine its sensitivity to the most influential features.
KNN showed improvement over Naive Bayes, with an accuracy of 69.54%, indicating its efficacy in capturing more complex patterns in the data.

```
kknn    0    1
   0 8923 3069
   1 1266  973
> knn_error_rate=sum(fit!=test$US)/length(test$US)
> print(knn_error_rate)
[1] 0.3046167
> accuracy <- (1-knn_error_rate)*100
> accuracy
[1] 69.53833
>
```

3.2.3. Random Forest Classifier:

Random Forest was chosen for its robustness and accuracy in classification tasks, particularly effective in reducing overfitting while handling high-dimensional data. Initially applied to a subset of the data, the Random Forest model was then scaled to the entire dataset, involving parameter tuning such as the number and depth of trees to balance bias and variance effectively. Random Forest outperformed the other models, achieving an accuracy of 72.99%. Its ability to provide insights into feature importance was crucial for understanding which variables most strongly predict booking destinations.



3.3. Model Evaluation:

Each model underwent rigorous testing and validation using cross-validation techniques and holdout sets to ensure generalization beyond the training data. This involved calculating accuracy, precision, recall, and F1-score for each model, as well as visually assessing performance using confusion matrices. Feature importance analysis, particularly with the Random Forest model, provided insights into the most influential predictors of a user's booking destination.

MODELS	ACCURACIES
NAIVE BAYES	68.07992
KNN	69.53833
RANDOM FOREST	72.98855

4. DISCUSSION

The project successfully predicted new users' first booking destinations on Airbnb with high accuracy (72.99%), primarily leveraging the Random Forest model. This model not only achieved impressive predictive performance but also provided valuable insights into the most influential features driving user behaviors.

- **Gender Disparity**: The analysis revealed a notable gender disparity, with a higher proportion of female users booking accommodations on Airbnb compared to males. Understanding this trend can inform targeted marketing strategies and personalized content offerings to better cater to the preferences of female users.

- **Age Demographics**: The age distribution of users indicated that the majority fell within the 30-40 age range. This demographic insight is crucial for tailoring accommodation

options and travel experiences to align with the preferences and expectations of users in this age group.

- **Signup Platforms**: The preference for web browsers (such as Safari and Chrome) as the primary signup platform highlights the importance of maintaining a user-friendly and efficient online interface. Optimizing the web experience can enhance user engagement and streamline the booking process.

4.1. Achievements and Implications:

The project successfully translated raw data into actionable insights, demonstrating the efficacy of data analytics in informing strategic decision-making. By understanding the factors influencing new users' booking decisions, Airbnb can enhance the overall user experience by offering more personalized content and recommendations.

- **Enhanced User Experience**: Personalized content and tailored recommendations can enhance user satisfaction and increase booking rates, ultimately improving the overall user experience on the Airbnb platform.

- **Support for Business Growth**: The analytical models developed in this project provide a foundation for refining Airbnb's strategies to support sustainable growth and user engagement. By leveraging data-driven insights, Airbnb can adapt to evolving user preferences and market dynamics more effectively.

4.2. Future Directions:

While the project has provided valuable insights into user booking behavior, there are opportunities for further exploration and improvement:

- **Model Optimization**: Future research could explore more complex algorithms or neural network architectures to capture subtleties in the data better and further improve predictive performance.

- **Real-Time Data Analysis**: Implementing the developed models in a real-time data processing environment would enable Airbnb to offer dynamic recommendations based on up-to-date user interactions, enhancing the responsiveness and relevance of the platform.

4.3. Limitations and Potential Challenges:

- **Data Limitations**: The study acknowledges potential limitations in the dataset, such as missing or incomplete data, which may have affected the accuracy and generalizability of the results. Addressing these data quality issues is essential for ensuring the reliability and robustness of future analyses.

- **Scope of Analysis**: The analysis focused primarily on predicting new users' first booking destinations within the existing dataset. Future research could explore additional factors, such as user reviews, preferences, or external market trends, to provide a more comprehensive understanding of user behavior and preferences on Airbnb.

References:

1. Le, N., & Naveed, F. (2020). Predicting First Booking Destinations on Airbnb: A Machine Learning Approach. Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), 4594–4603. https://doi.org/10.1109/bigdata50022.2020.9378032
2. Blecharczyk, N., Gebbia, J., & Chesky, B. (2017). How Airbnb designs for trust. TED Talks. Retrieved from https://www.ted.com/talks/nathan_blecharczyk_how_airbnb_designs_for_trust/transcript
3. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686.
4. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
5. Dowle, M., & Srinivasan, A. (2021). data.table: Extension of **data.frame**.
6. Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22.
7. Wickham, H. (2011). The split-apply-combine strategy for data analysis. Journal of Statistical Software, 40(1), 1-29.
8. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.
9. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
10. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.
11. Website and Online Resources:

a. Airbnb. (2020). About Us. Retrieved from Airbnb Website

b. Medium Article on Data Preprocessing Techniques in R. Retrieved from Medium

12. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.