Travel Package Prediction

RASHI MAKADIA(229309045), ARNAV GUPTA (229309034),
KRITI GUPTA (229309067), HARSHITA BATTA (229309044)

MANIPAL UNIVERSITY JAIPUR

I. INTRODUCTION

In today's age, the travel industry is at the forefront of innovation and constantly looking for new ways to improve the customer's travel experience. One way to do so would be creating a travel package predictor, which could predict which package a customer would end up buying depending on the attributes of the customer. A travel package predictor can help companies in increasing sales by predicting which customer is more likely to buy which package, meaning showing the right ad to the right person, which can lead to more sales. Instead of a generic approach, travel companies can use predictors to recommend packages that better suit a customer's interests and budget. This personalization can lead to a more positive customer experience.

By analysing past data, companies can determine which package was popular with which customer group, and can potentially create new packages that would appeal to certain customers more than the pre existing ones. This can also help with reducing costs for both the company, as well as making trips budget friendly for customers.

The dataset we're using is taken from kaggle. It has been worked on previously by 2 people, who used ensemble techniques such as Decision Trees, Random Forest, Boosting, Bagging etc. for predictions.

Even though just like them, we're also using ensemble techniques, by incorporating dataset balancing techniques and hyperparameter tuning for models, we aim to achieve more accurate predictions.

II. METHODOLOGY

a. DATA DICTIONARY

Customer Details

This section contains features that will help us understand the demographic characteristics and history of the customers.

- **CustomerID:** Unique identifier for each customer
- **ProdTaken:** Binary variable (0 or 1) representing whether the customer has purchased a travel package or not
- **Age:** The customer's age in years
- **TypeofContact:** How the customer was contacted (Company Invited or Self Inquiry)
- **CityTier:** Categorization of cities based on development, population, and lifestyle
- **Occupation:** Customer's profession
- **Gender:** Customer's gender
- **NumberofPersonVisiting:** Total number of people planning to travel with the customer
- **PreferredPropertyStar:** Customer's preferred star rating for accommodation
- **MaritalStatus:** Marital status of the customer
- **NumberOfTrips:** Number of trips the customer has taken in the past year
- **Passport:** Binary variable (0 or 1) indicating whether the customer has a passport
- **OwnCar:** Binary variable (0 or 1) representing whether the customer owns a car
- **NumberOfChildrenVisiting:** Total number of children under the age of 5 planning to travel with the customer
- **Designation:** Customer's job title in their current organisation (if applicable)
- **MonthlyIncome:** Customer's monthly income (if provided)

Customer Interaction Data

This section includes information about the customer's interaction with the travel agency or company.

- **PitchSatisfactionScore:** Sales pitch satisfaction score
- **ProductPitched:** Specific travel package presented by the salesperson
- **NumberOfFollowups:** Number of times the salesperson contacted the customer after the pitch
- **DurationOfPitch:** Time spent by the salesperson presenting the travel package to the customer

b. EXPLORATORY DATA ANALYSIS

1. **Univariate Analysis -** This analysis determines the central tendency (mean, median, mode) and dispersion (standard deviation) of quantitative data by examining the distribution of each variable and a group of data.

2. **Bivariate Analysis -** This stage examines the relationship between pairs of variables and identifies relationships or relationships that may influence

purchasing decisions. Techniques such as scatter plots and correlation coefficients will be used.

3. **Multivariate Analysis -** This analysis examines the relationship between multiple variables simultaneously to better understand the factors that influence purchasing behaviour.

c. DATA PREPROCESSING TECHNIQUES

The dataset must be strictly preprocessed to provide accurate results. Pre processing has been divided into several steps:

1. **Outlier Detection -** Outliers are data points that differ from the distribution. Techniques such as visualisation (box plots) or statistical methods (interquartile range - IQR) can be used to identify potential outliers. Depending on the analysis, the output can be extracted, queued (only one value), or processed using a well-defined analysis method.

2. **Handling Missing Values -** Missing values in the dataset may negatively affect the learning model. Techniques such as mean, median, or mode imputation are used to resolve missing values and replace them with appropriate values based on the mode and distribution of the variable.

3. **Converting categorical data to numeric data -** Machine learning algorithms typically work best with numerical data. Categorical data needs to be converted into a format the algorithm can understand. This conversion process, often called encoding, allows the model to recognize the relationships between different categories and use that information for predictions. Without encoding, the model wouldn't be able to distinguish between "high" and "low" income, treating them as completely separate entities instead of ordered values.

4. **PCA -** It is a dimensionality reduction technique that can be used when a data set has many features. In the case of our dataset, PCA was not used as dimensionality reduction was changing the accuracy of the model drastically.

5. **Data Balancing -** Data balancing is a technique used in machine learning to address the issue of imbalanced datasets. A dataset is considered imbalanced when one class (or category) has significantly more examples than others. This can cause problems for machine learning models because they tend to become biased towards the majority class.

   SMOTE ENN is a combination of two techniques used for imbalanced data:

SMOTE (Synthetic Minority Oversampling Technique): This technique creates synthetic data points for the minority class to address class imbalance.

ENN (Edited Nearest Neighbors): This technique is an undersampling technique that removes noisy data points from both the majority and minority classes.

It was implemented through the imbalanced-learn library to address the issue of imbalanced data.

d. MODELLING

We examined various machine learning algorithms for creating powerful predictive models.

1. **Decision Tree -** A decision tree is a machine learning algorithm used for classification and regression tasks. It creates a tree by recursively splitting the data until a halting condition is satisfied, dividing a dataset into subsets according to input features. After determining which feature is the best, the algorithm generates a leaf node that symbolizes the choice. Both continuous and categorical variables can be handled by decision trees.

2. **Random Forest -** Random Forest is an ensemble learning method that uses multiple decision trees to make predictions. It combines the average prediction in regression tasks and the mode of class predictions in classification problems to make predictions. This reduces overfitting and improves generalization performance by reducing variance. Through the random selection of a subset of characteristics at each node, it can handle enormous feature sets in high-dimensional datasets.

3. **Bagging -** Bagging is an ensemble learning technique that improves the stability and accuracy of machine learning algorithms, particularly decision trees. It generates multiple subsets of the original dataset, resamples it, and trains a base model on each subset. This reduces variance by averaging or combining the predictions of multiple models trained on different data subsets. Bagging also reduces sensitivity to small variations in training data, leading to improved predictive performance compared to a single model.

4. **AdaBoost -** AdaBoost is an ensemble learning method that uses multiple weak learners, like decision trees, to create a strong classifier. It focuses on difficult instances and gives more weight to those misclassified by previous weak learners. The process is iterative, with each weak learner trained sequentially. The weights of misclassified instances are adjusted after each iteration for perfect

classification. AdaBoost combines predictions using a weighted majority vote or weighted average, and the final prediction is obtained by aggregating all weak learners' predictions.

5. **Gradient Boosting -** Gradient Boosting is an ensemble learning technique that builds a strong learner by combining the predictions of multiple weak learners, like decision trees. It builds the ensemble sequentially, with each weak learner correcting previous errors. The process starts with an initial prediction and iteratively fits a weak learner to the residuals of the current ensemble's predictions. The weak learner is trained to capture errors from previous iterations. The ensemble is updated by adding a scaled version of the weak learner to minimize the loss function.

6. **XGBoost -** XGBoost is a scalable and efficient implementation of the Gradient Boosting framework for supervised learning tasks like classification, regression, and ranking. It improves performance and scalability by handling complex datasets with large features and instances. XGBoost uses techniques like approximate tree learning, parallel computing, and cache-aware access to optimize performance. Regularization techniques like L1 and L2 can be tuned during training to balance model complexity and performance.

e. HYPERPARAMETER TUNING AND CROSS VALIDATION

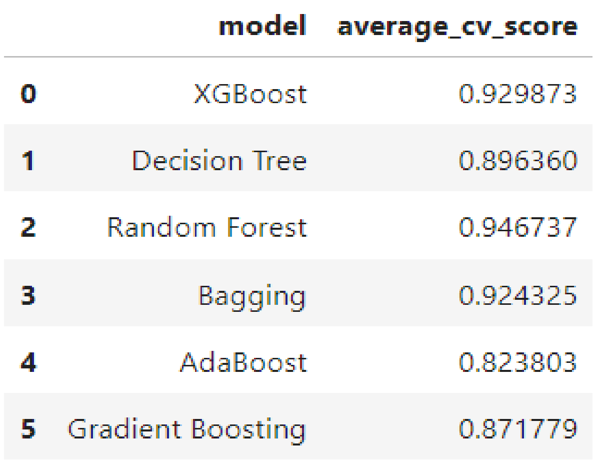
Hyperparameter tuning process involves selecting the best fit set of hyperparameters for a machine learning algorithm. Hyperparameters are parameters that are set before the learning process begins, such as the learning rate in gradient descent or the depth of a decision tree. Hyperparameter tuning is typically done through techniques like grid search, random search, or more advanced methods like Bayesian optimization.

Cross-validation is a technique used to assess how well a model will generalise to new, unseen data. The model is trained on some of the folds and then evaluated on the remaining fold(s). This process is repeated multiple times, with different subsets used for training and evaluation each time. In our model we have used 5 folds for cross-validation.

III. RESULTS

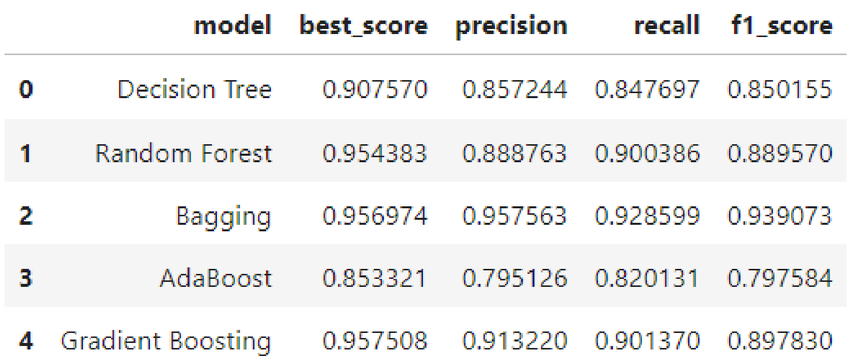
After carefully selecting and implementing all the algorithms and the appropriate EDA techniques that are suitable for our topic and dataset, we came to the conclusion about the group of people that should be targeted for a particular travel package listed by the company.

Before performing hyperparameter tuning, Random Forest was giving the highest accuracy.



	model	average_cv_score
0	XGBoost	0.929873
1	Decision Tree	0.896360
2	Random Forest	0.946737
3	Bagging	0.924325
4	AdaBoost	0.823803
5	Gradient Boosting	0.871779

After hyperparameter tuning, results from all the models were compared:

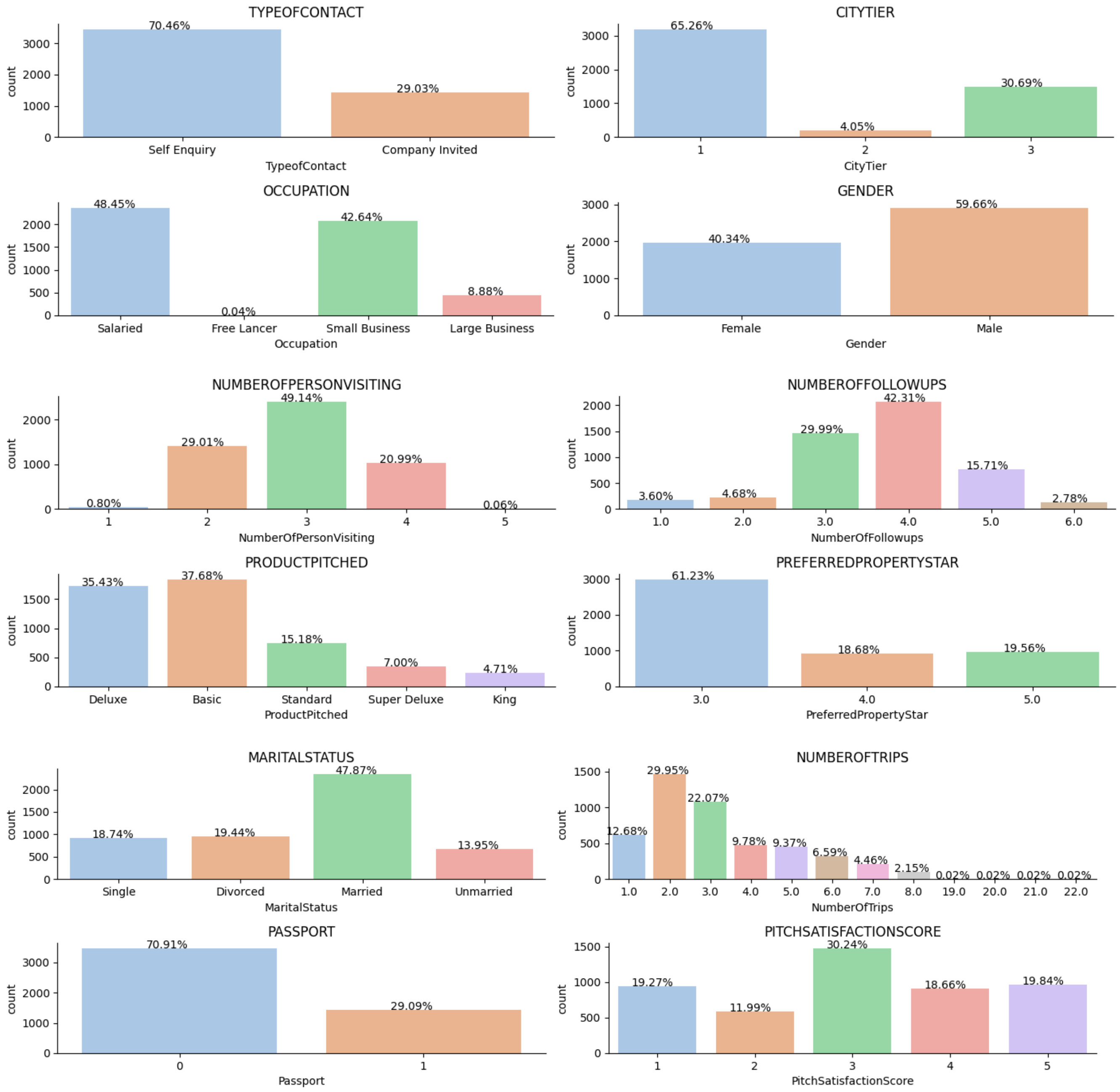


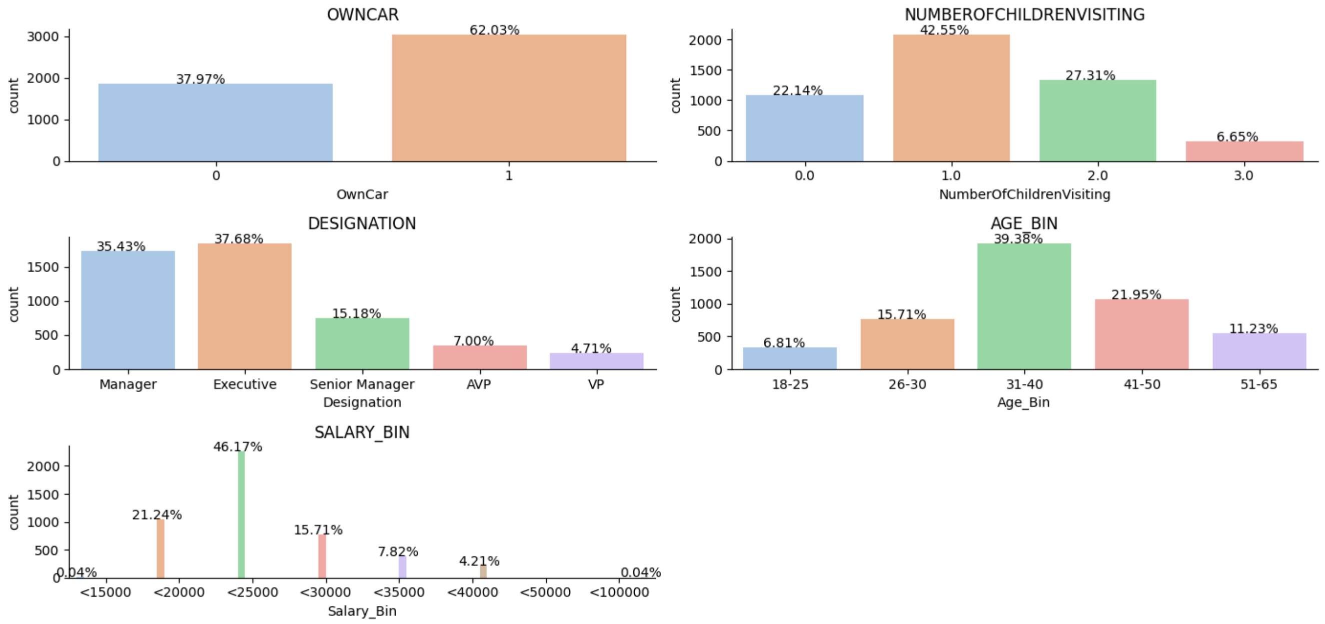
	model	best_score	precision	recall	f1_score
0	Decision Tree	0.907570	0.857244	0.847697	0.850155
1	Random Forest	0.954383	0.888763	0.900386	0.889570
2	Bagging	0.956974	0.957563	0.928599	0.939073
3	AdaBoost	0.853321	0.795126	0.820131	0.797584
4	Gradient Boosting	0.957508	0.913220	0.901370	0.897830

Gradient Boosting achieved highest accuracy after hyperparameter tuning.
Other metrics such as precision, recall and F1 score were also calculated for further insights.

IV. ANALYSIS

A. UNIVARIATE ANALYSIS


TYPEOFCONTACT
count
0
1000
2000
3000
70.46%
29.03%
Self Enquiry
Company Invited
TypeofContact
CITYTIER
count
0
1000
2000
3000
65.26%
4.05%
30.69%
1
2
3
CityTier
OCCUPATION
count
0
500
1000
1500
2000
48.45%
0.04%
42.64%
8.88%
Salaried
Free Lancer
Small Business
Large Business
Occupation
GENDER
count
0
1000
2000
3000
40.34%
59.66%
Female
Male
Gender
NUMBEROFPERSONVISITING
count
0
1000
2000
0.80%
29.01%
49.14%
20.99%
0.06%
1
2
3
4
5
NumberOfPersonVisiting
NUMBEROFFOLLOWUPS
count
0
500
1000
1500
2000
3.60%
4.68%
29.99%
42.31%
15.71%
2.78%
1.0
2.0
3.0
4.0
5.0
6.0
NumberOfFollowups
PRODUCTPITCHED
count
0
500
1000
1500
35.43%
37.68%
15.18%
7.00%
4.71%
Deluxe
Basic
Standard
Super Deluxe
King
ProductPitched
PREFERREDPROPERTYSTAR
count
0
1000
2000
3000
61.23%
18.68%
19.56%
3.0
4.0
5.0
PreferredPropertyStar
MARITALSTATUS
count
0
500
1000
1500
2000
18.74%
19.44%
47.87%
13.95%
Single
Divorced
Married
Unmarried
MaritalStatus
NUMBEROFTRIPS
count
0
500
1000
1500
12.68%
29.95%
22.07%
9.78%
9.37%
6.59%
4.46%
2.15%
0.02%
0.02%
0.02%
0.02%
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
19.0
20.0
21.0
22.0
NumberOfTrips
PASSPORT
count
0
1000
2000
3000
70.91%
29.09%
0
1
Passport
PITCHSATISFACTIONSCORE
count
0
500
1000
1500
19.27%
11.99%
30.24%
18.66%
19.84%
1
2
3
4
5
PitchSatisfactionScore

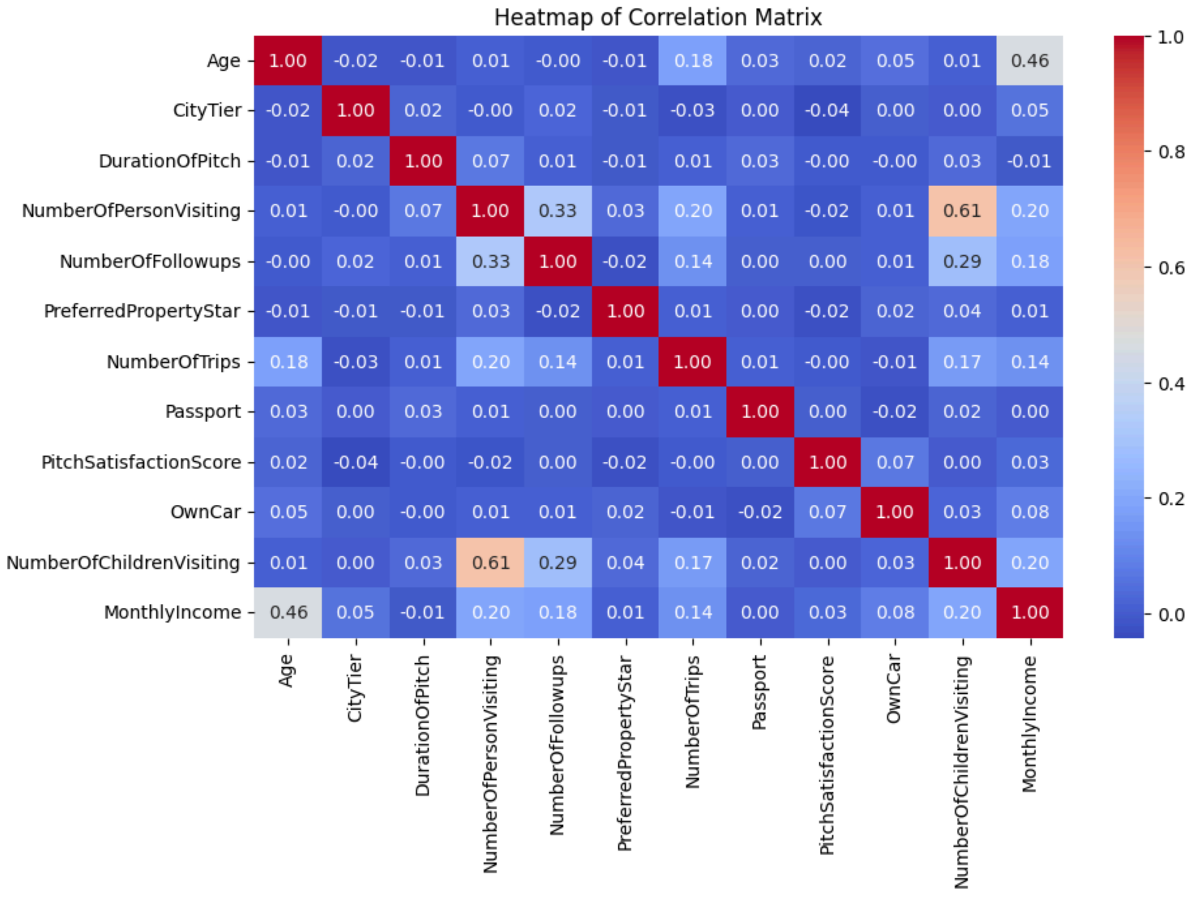


- Approximately 38% of customers hold executive positions, with managers following closely at 35%.
- About 18% of customers accepted the product offered during the last interaction.
- A majority, around 62%, of customers own a car.
- Nearly 29% of customers possess a passport.
- Around 65% of customers reside in Tier 1 cities.
- Approximately 61% of customers prefer 3-star accommodations.
- Nearly half, about 48%, of customers are married.
- Basic package pitches were made to roughly 38% of customers, while Deluxe packages were pitched to 35%.
- Approximately 60% of customers are male.
- Nearly half, about 49%, of customers are salaried individuals.
- A significant 70.5% of customers initiated inquiries for the packages themselves.
- Most customers travelled with three people.
- The majority of customers take two trips per year, although there are some outliers, such as 22 trips.
- Most customers travelled with only one child.
- Typically, customers were followed up four times.
- The majority of customer monthly incomes fall within the 20,000-25,000 range, with most falling between 15,000-30,000 monthly.
- About 35% of customers are in the 31-40 age group, with the majority falling within the 26-50 age range.

B. BIVARIATE AND MULTIVARIATE ANALYSIS

1. HEATMAP

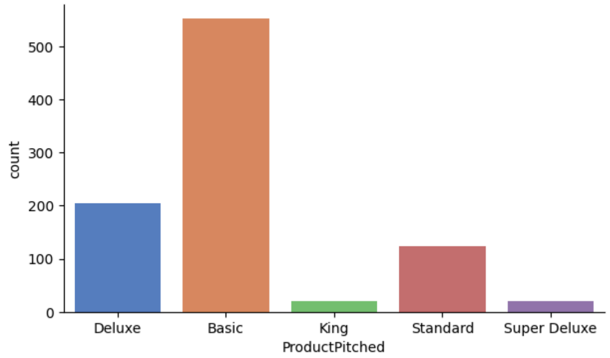
Implementing multivariate analysis on our dataset, we generated a heatmap to visualise the correlations between different variables. An intriguing finding from our analysis is the strong correlation between the number of children visiting and the total number of persons visiting. This observation suggests a potential relationship between family size and the presence of children, which could have significant implications for designing and marketing travel packages targeted at families. Understanding these correlations enables us to tailor our offerings more effectively to meet the needs and preferences of our target audience.

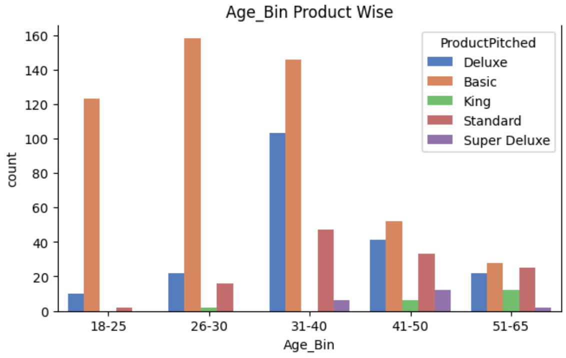


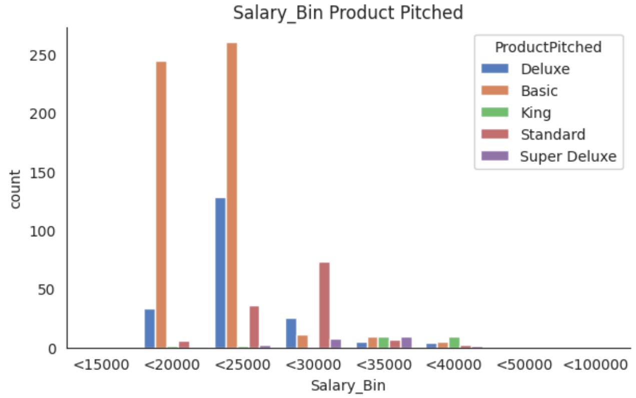
- Monthly income and age display a slight correlation, implying that age's influence on income is limited.

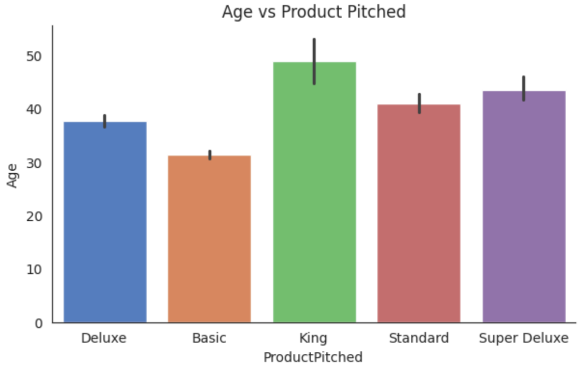
- The number of children visiting correlates as anticipated with the total number of visitors.
- There's a notably weak correlation between the number of follow-ups and the initial count of visitors, indicating a loose connection between follow-up needs and initial visitation numbers.

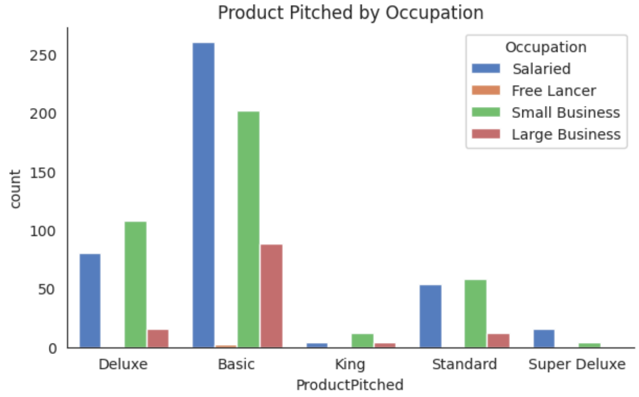
2. CUSTOMER PROFILE BY PRODUCT TYPE

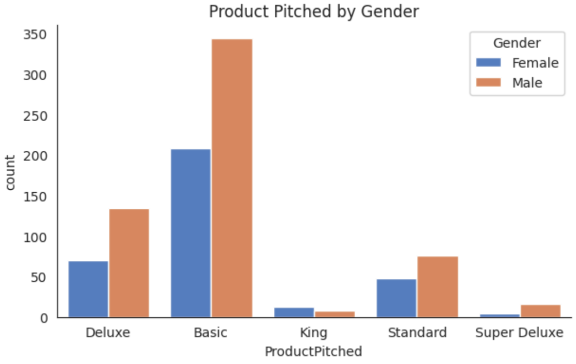


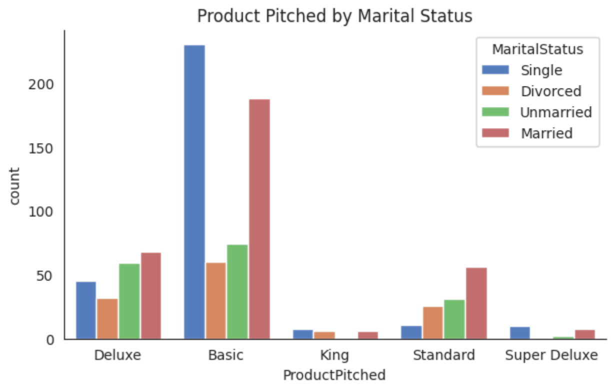


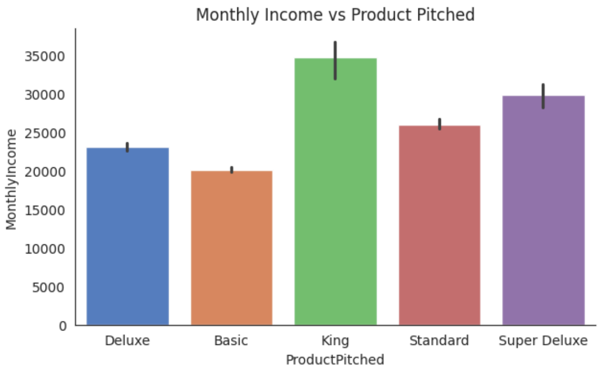


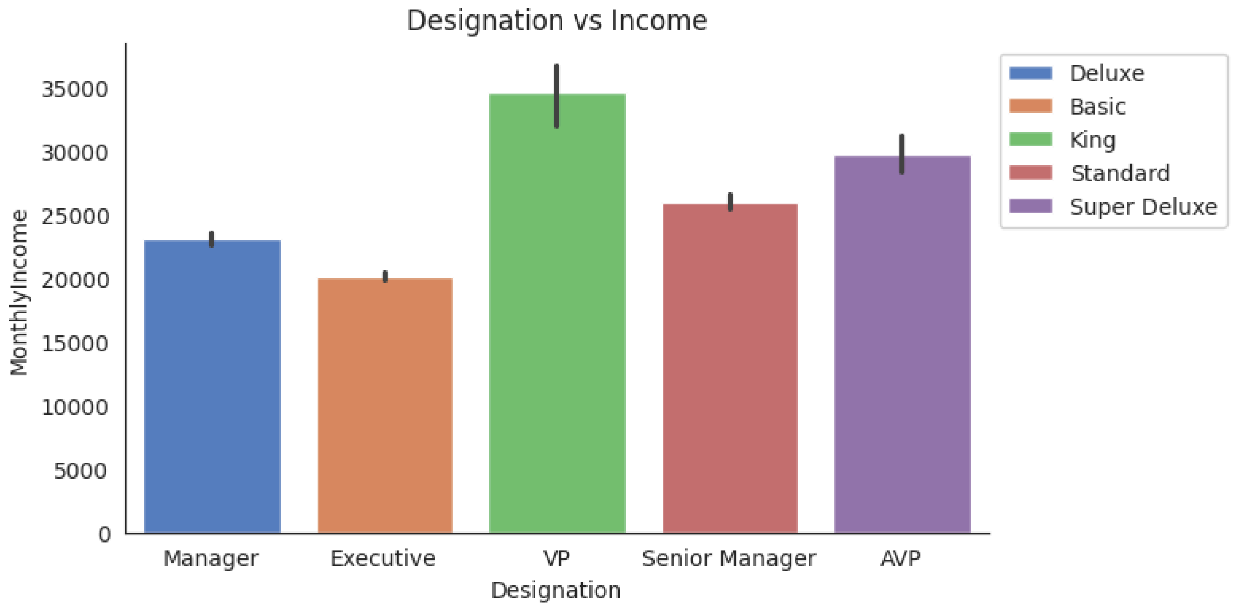


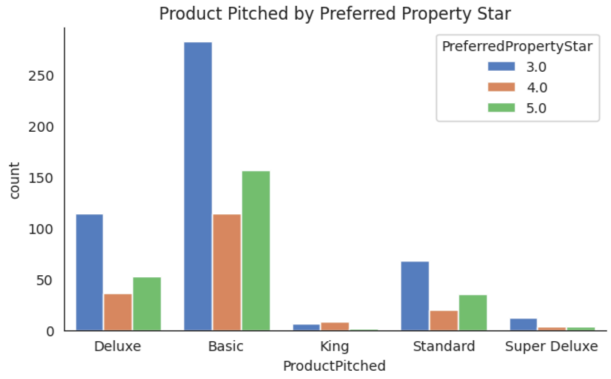


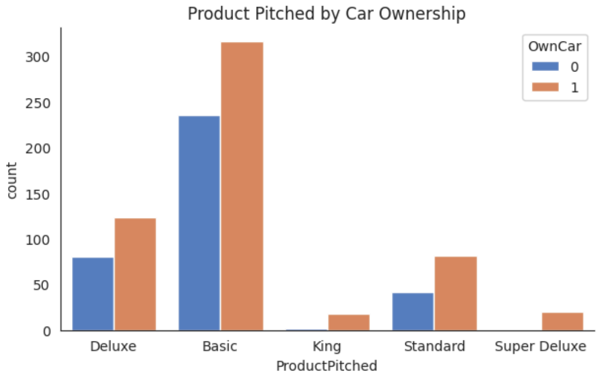












Based on the analysis of customer profiles according to the product pitched and purchased, distinct trends emerge:

Basic Package: Customers with a monthly income below 25,000, aged between 26-30, holding executive positions, residing in tier 1 cities, and primarily single males, show a preference for this package. Interestingly, married individuals also lean towards the basic package.

Deluxe Package: Customers with a monthly income below 25,000, aged between 31-40, holding managerial positions, often married, and residing in tier 3 cities tend to opt for the deluxe package. This preference extends to divorced customers and those from tier 1 cities.

King Package: Individuals with a monthly income ranging from 30,000 to 35,000, aged between 51-60, holding VP positions, residing in tier 1 cities, and predominantly single females engaged in small businesses are inclined towards the king package. Notably, females are more likely to purchase this package compared to men.

SuperDeluxe Package: Customers with a monthly income below 35,000, aged between 41-50, holding AVP positions, residing in tier 3 cities, and primarily single males working in salaried occupations show a preference for the SuperDeluxe package. A significant proportion of them are company-invited.

Standard Package: Individuals with a monthly income below 30,000, aged between 31-40, holding senior managerial positions, often married, and from tier 3 cities, typically engaged in small businesses, prefer the standard package. This preference is frequently observed among self-inquired customers.

These distinct customer profiles provide valuable insights for targeted marketing strategies, enabling companies to tailor their offerings and outreach efforts to better match the preferences and characteristics of their target audience, ultimately enhancing customer engagement and conversion rates.

C. PERFORMANCE METRICS

1. **Precision:**
   Precision is a statistical metric used to assess the accuracy of a model's predictions by determining the proportion of true positive predictions among all positive predictions. Precision is calculated using the following formula:
   Precision = True Positives / (True Positives + False Positives)
   Precision measures the accuracy of a model's positive identifications, indicating its ability to avoid labeling negative instances as positive.
   Here’s what each component of the formula represents:
   - True Positives (TP): The number of correctly predicted positive instances.

- False Positives (FP): The number of negative instances incorrectly predicted as positive by the model.

A high precision score indicates a model's ability to reduce false positive predictions, indicating better identification of positive instances. Conversely, a low precision score indicates incorrect labeling of negative instances, resulting in higher false positives. Precision is crucial in medical diagnosis.

2. **Recall:**
   Recall, also known as sensitivity or true positive rate, is a crucial metric in statistics and machine learning to assess the accuracy of classification models.
   Recall is calculated using the following formula:
   Recall = True Positives / (True Positives + False Negatives)
   Recall measures the model's accuracy in identifying positive instances in the dataset, ensuring all relevant instances are captured without false negatives.
   Here's what each component of the formula represents:
   - True Positives (TP): The number of correctly predicted positive instances.
   - False Negatives (FN): The number of positive instances incorrectly predicted as negative by the model.

   A high recall score indicates the model captures most positive instances in the dataset, reducing false negatives. A low recall score indicates significant missing instances, increasing false negatives. Recall is crucial in scenarios like medical diagnosis, as it indicates the proportion of correctly identified cases.

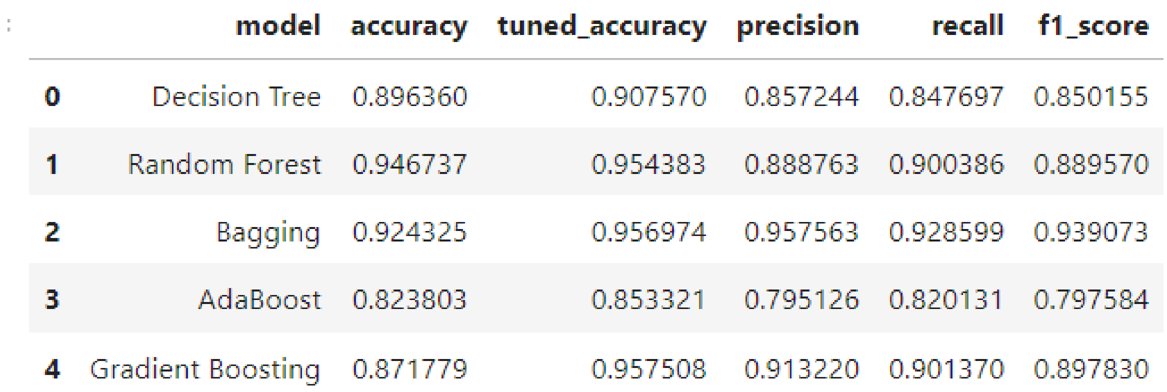
3. **F1 score:**
   The F1 score is a metric that combines precision and recall, offering a balanced evaluation of a model's performance in binary classification tasks. It is the harmonic mean of precision and recall and is calculated using the following formula:
   F1 = 2 * (Precision * Recall) / (Precision + Recall)
   The F1 score evaluates models considering both false positives and false negatives, assisting in assessing models with imbalances between positive and negative classes or when precision and recall are crucial.
   The F1 score ranges from 0 to 1, where:
   - A high F1 score (close to 1) indicates both high precision and high recall, meaning the model is performing well in terms of both minimizing false positives and false negatives.
   - A low F1 score (close to 0) indicates poor performance in either precision or recall, or both.

The F1 score is useful for comparing models' performance without favoring precision or recall alone, like in spam email detection systems. It measures a model's ability to achieve both objectives simultaneously.

V. CONCLUSION

With this project, we explored the applications of ensemble learning techniques, namely Random Forest, Bagging, AdaBoost, Gradient Boosting, XGBoost, along with Decision Trees in predicting travel package preferences. We have demonstrated the necessity of predictive models in the travel industry, given the increasing demand for personalised travel experiences.

We analyse our model in terms of accuracy, precision, recall and F1 score. By using grid search to find optimal parameters and using cross validation while training the model, by comparing results from the different algorithms, we're able to determine which algorithm performs best on our dataset.



	model	accuracy	tuned_accuracy	precision	recall	f1_score
0	Decision Tree	0.896360	0.907570	0.857244	0.847697	0.850155
1	Random Forest	0.946737	0.954383	0.888763	0.900386	0.889570
2	Bagging	0.924325	0.956974	0.957563	0.928599	0.939073
3	AdaBoost	0.823803	0.853321	0.795126	0.820131	0.797584
4	Gradient Boosting	0.871779	0.957508	0.913220	0.901370	0.897830

While performing GridSearchCV, due to computational limitations, we weren't able to find the optimal parameters for XGBoost, hence, omitting it in the final comparison. While analysing the dataset, we noticed that it was imbalanced and undersampling the dataset would drop important information, so we used SMOTE ENN for balancing the dataset while preserving important information.

Looking ahead, future research directions may involve exploring additional ensemble learning techniques, incorporating novel data sources such as real-time social media data or geolocation information, and investigating the application of our model in diverse cultural and geographical contexts. By continuously pushing the boundaries of innovation in travel prediction, we can further advance the state-of-the-art and pave the way for a more personalised and fulfilling travel experience for all.

VI. REFERENCES

1. https://www.kaggle.com/datasets/sanamps/tourpackageprediction/data
2. Brown, G. (2010). Ensemble Learning.. Encyclopedia of machine learning, 312, 15–19.
3. Mishra, S. (2017). Handling imbalanced data: SMOTE vs. random undersampling. Int. Res. J. Eng. Technol, 4(8), 317–320.