Dean&Francis

ISSN 2959-6130

Predictive Business Analytics using Different Predicting Model, Comparison of Decision Tree and Random Forest

Chen Chen[1,*]

[1]*School of Economics, University of Edinburgh, Edinburgh, United Kingdom*
**corresponding author: c.chen-111@sms.ed.ac.uk*

Abstract:

Predictive Analytics is a crucial branch for data-driving analysis in business, and could be used in cutomer behavior prediction and evaluation. The selection of an appropriate model involves a trade-off between prediction accuracy, interpretability and computational costs. This study aims to empirically compare the predictive performance of Decision Tree (DT) and Random Forest (RF) in identifying potential customer type and bahavior for a new travel package, addressing a challenge of model selection in real-business context. Using a customer dataset from a tourism company ("Visit with us"), it implemented and tuned both DT and RF models. Model performance was evaluated on key metrics including accuracy, precision, recall, and F1-score. The analysis found that tuned Random Forest Model is found to be the superior model with the highest test accuracy and F1 score, indicating a balance between precision and recall. While tuned Decision Tree Model ranked at the second. The study confirms the established superiority of ensemble methods like Random Forest for prediction tasks but provides a nauced business insight: the choice between DT and RF might depend on the strategic goal—maximizing customer reach versus optimizing marketing efficiency. The findings offer actionable strategies for targeted marketing and demonstrate the significant value of predictive analytics in formulating business strategy.

Keywords: Predictive analysis, machine learning, decision tree, random forest, customer behavior.

1. Introduction

As one of the quantitative methods, Business Analytics (BA) derives meaningful insights from datasets using statistical techniques. And BA could be applied across various sectors, such as stock markets, banking, medicine, and the retail industry, for forecasting and supporting data-driven decision making. Recent-

ly, the focus on BA has shifted to Predictive Analytics (PA).

Predictive Analytics is a branch that employs input data, statistical methods and machine learning models to forecast future trends and the probability of some specific outcomes [1]. There are numerous studies on PA topic, "Kumar and Garg, 2018" summarize the process and techniques used in predictive analytics, such as decision tree, regression model, artificial neural network and other key techniques. This paper provides key implementations of using predictive analytics in banking and financial services for minimizing credit risks and attracting some valuable potential customers, helping the retail industry to predict customers' behaviors and the reaction of a new product in the market, and predicting drug design to improve supply chain management [2]. In this paper, the decision tree and random forest are mainly used in supervised models with the goal of the algorithm to obtain a classifier by learning from a training dataset. This classifier can be used for classification and prediction in testing samples.

Numerous studies are using Decision Tree in predicting under different scenarios. Lee et al. summarized the methodology of Decision Tree (DT) based on 20 research papers in various fields and applications, showing that DT has significant potential as PA tool in the supply chain, as it is easy to interpret without any deep understanding of statistical methods [3]. While Gepp et al. extend the study done in Frydman et al. and include Cart and See5 decision trees in business failure prediction in the US manufacturing and retail industry using the Computer Statistics (COMPUSTAT) database, and concluded that DT techniques are superior in predictive and classificatory performance in business failure prediction [4]. The random forest algorithm is also used in prediction. Valecha et al. used this methodology to predict and classify customer buying behavior using this kind of algorithm, achieving an accuracy rate of 94%. Their study found that customer behavior in a competitive market fluctuates frequently and depends on several factors, including environmental, organizational, individual, and interpersonal factors [5].

In this paper, decision tree and random forest algorithms will be used to predict future customer behavior based on current data from a tourist company. As incomes grow worldwide with economic development, the demand for leisure has become an inevitable part of consumer spending. Therefore, it is crucial to understand the consumer mindset to meet the demands of both existing and potential customers in the market. However, preferences for goods and services can vary significantly among different customers. Hence, analytical methods can be employed to explore historical and current customer data to understand their needs and summarize behavioral trends. This can be a challenging task, as the consumer's mind is influenced not only by personal factors but also significantly by external environmental factors. Therefore, predicting actual customer behavior requires more than just identifying determining features; it serves as a reference for informing future marketing strategies.

2. Model Introduction

2.1 . Decision Tree

A common technique used in data mining involves building classifiers. And the algorithms of classification are designed to handle larger datasets, derive insights from a training set with known labels and classify new data subsequently. This paper mainly focuses on the decision tree algorithm and random forest algorithm with application on predicting tourist package customers' future behavior.

A Decision Tree is a tree-like model contains root node, branches and leaf nodes. And it helps to solve the classification problem using nominal data that has no natural ordering or quantifiable similarity. It classifies information by asking sequences of questions such as "yes/no", "true/false" or "value (property), does the value belong to the set", where the next question asked is dependent on the previous question [6].

The standard decision tree process, as used in this paper, like C4.5 and Cart grows from the root node at the top down using a depth-first, divide-and-conquer strategy. The algorithm begins by selecting an attribute for the root node based on a splitting criterion such as information gain or Gini Index, and creating branches for its possible outcomes. This partitions the training data into distinct subsets corresponding to each branch. Then the method repeats the splitting process for each subsequent branch, using only the subset of instances that proceed down that path. And the tree is expanded traditionally from left to right. Recursion stops at any node that shares the same class label, which is known as a pure label. And the tree is built until all nodes are pure and followed by pruning to mitigate the overfitting issue [7].

The pruning in a decision tree includes pre-pruning and post-pruning. The pre-pruning solution occurs when the decision to stop splitting at a node is made without considering the potential for highly informative splits further down the tree. Therefore, this method biases the tree towards making its most important decisions only near the root, potentially missing more complex patterns and leading to suboptimal accuracy. Post-pruning serves as an alternative to stopping splitting to mitigate the "horizon effect." The principal strategy involves first growing the tree completely until the leaf nodes achieve minimal impurity. Then, the algorithm works backward, considering pairs of leaf nodes with a common parent for elimination. A pair of leaves is merged if this action results in only a small satisfactory increase in overall impurity for the elimina-

ISSN 2959-6130

tion criterion. In our application setting, the pre-pruning method is used to limit the maximum depth of the tree [8].

2.2 . Random Forest

The decision tree is prone to overfitting. But if it do prune or limit the maximum depth of the tree, there is a loss in classification accuracy and precision. Ho introduced a method for growing multiple decision trees using randomization, where trees are built from randomly selected feature subspaces, with each tree often achieving 100% accuracy on the training data [9]. This concept was later extended and formally presented as the Random Forest algorithm by Breiman, who defined it as an ensemble of tree-structured classifiers. It is defined that classifiers are independent identically distributed random vectors that vote for the most popular class for classification tasks or take the average prediction for these trees. The process begins by randomly selecting an equal-sized portion of data from the original training dataset. Additionally, a random subset of features is chosen from building each decision tree fully and randomization is used to reduce correlation among different decision trees [10].

2.3 . Comparison of Decision Tree and Random Forest

One primary advantage is the interpretability of decision tree models as the logic behind predictions is transparent and can be easily summarized from the characteristics at each leaf node. Meanwhile, this machine learning method does not need any data preparation, such as normalization of the whole dataset. However, general disadvantages are using this kind of model. Firstly, the most common one is that decision trees are prone to overfitting, especially for small datasets, so even a small change in data might lead to different trees. Also, it is not an ideal model for a large dataset with too many features due to computational cost and it might create a biased tree with a dominating feature.

For the random forest algorithm, one advantage is the versatility, as this approach is applicable for both regression and classification issues. Additionally, this algorithm is advantageous because it produces superior results without requiring hyperparameter tuning. Furthermore, its operations are highly transparent and easily understandable [11]. An individual decision tree has the advantages of interpretability, while a random forest, as a combined trees model, loses interpretability. In exchange, random forest always has a better performance in prediction.

3. Methodology

The notebook from Kaggle.com from a tourist company named “Visit with us”, and this company wants to use this business analytics model to expand its customer base. Based on the data of the previous year, the company plans to launch a new product called “Wellness Tourism Package” and use models to predict three problems: which kind of customers are more likely to purchase the newly introduced travel package, which feature is most significant in determining purchasing and which segment of customers should be targeted more in the marketing strategy. In the end, offering business recommendations for the Marketing Department by predicting the potential customer who is going to buy this new package.

3.1 . Dataset Selection

3.1.1 . Statistical Distribution of Raw Dataset

Firstly, the dataset is collected from Kaggle.com on travel package purchase prediction, which is named ‘tour_package.csv’. The raw dataset contains 4888 rows (4888 customer history data), and 20 feature columns. The features are divided into numerical and categorical values, where the description of the features is shown below. The dependent variable in this notebook is ProdTaken, which means whether the customer has purchased a package as shown in Table 1 and Table 2.

Table 1: The summary of numerical features

Variable	Meaning	Count	Mean	std	Min	25%	50%	75%	Max
Age	Age of Customer	4662	37.62	9.32	18	31	36	44	61
Duaration On Pitch	Duaration of the pirch by a salesperson to the customer	4637	15.49	8.52	5	9	13	20	127
Monthly Income	Gross monthly income of the customer	4655	23619.85	5380.70	1000	20346	22347	25571	98678
Number of Trips	Average number of trips in a year by customer	4748	3.24	1.85	1	2	3	4	22

Number of Followups	Total Number of follow-ups has been done by the salesperson after the sales pitch	4843	3.71	1.00	1	3	4	4	6
Preferred Property Star	Preferred hotel property rating by customer	4862	3.58	0.80	3	3	3	4	5
Number Of Children Visiting	Total number of children with age less than 5 planning to take the trip with the customer	4822	1.19	0.86	0	1	1	2	3

Table 2: The summary of categorical features

Column	Meaning	Populated	Zero Values	Unique Values	Most Common	% Populated
Designation	Designation of the customer in the current organization	4888	0	5	Exceutive	100
ProdTaken	Whether the customer has purchased a package or not (0: No, 1: Yes)	4888	3968	2	0	100
OwnCar	Whether the customers own a car or not (0: No, 1: Yes)	4888	1856	2	1	100
Passport	The customer has a passport or not (0: No, 1: Yes)	4888	3466	2	0	100
CityTier	City tier depends on the development of a city, population, facilities, and living standards. The categories are ordered i.e. Tier 1 > Tier 2 > Tier3	4888	0	3	1	100
MaritalStatus	Marital status of customer(Married,Divorced,Single,Unmarried)	4888	0	4	Married	100
ProductPitched	Product pitched by the salesperson(Basic,Dekuxe,-Standard,SuperDeluxe, King)	4888	0	5	Basic	100
Gender	Gender: Gender of customer(Male, Female, Fe Male)	4888	0	3	Male	100
Occupation	Occupation of customer(Salaried, Small Business, Large Business, Free Lancer)	4888	0	4	Salaried	100
TypeofContact	How customer was contacted (Company Invited or Self Enquiry)	4888	0	2	Seld Enquiry	100

3.1.2 . Data Preprocessing/ Data Cleaning: Dealing with Missing Values

For categorical features, it first drop the Customer ID as it does not provide any information about customer details. Then, fix the Fe male group into Female to prevent misinterpretation. To explore any pattern in age and income, they are converted into age bins and income bins.

When addressing missing values in a dataset, there are four primary approaches for imputing or filling in the missing entries. For the "Type of Contact" column, missing values were imputed using the mode (the most frequently occurring value). For numerical features like "Number of Follow-ups," "Preferred Property Star," and "Duration of Pitch," the median value calculated within specific groups was used. For the remaining features—such as "Number of Trips", "Number of Children Visiting", "Age", "Duration of Pitch" and "monthly income"—a more granular approach was taken. Missing values were filled based on their relationship with other features like Gender, Designation, and Marital Status to ensure a more accurate and context-aware imputation.

3.1.3 . Customer Profile Based on Product Type

Before constructing the decision tree and random forest algorithm, the Exploratory Data Analysis computes the customer profile based on different types of the products which are Basic, Deluxe, King, SuperDeluxe and Stan-

ISSN 2959-6130

dard. This analysis, it analyzes the characteristics of customers based on their age range, monthly income, designation with city tier, gender and marital status. The results are summarized in the Table 3.

Table 3: The summary of customer profile

Product Type	Monthly income	Age	Designation & City tier	Martial Status	Occupation&Gender	Contact Information
Basic Package	<25000	26-30	Executive belong to City tier 1	Married	Salaried Males	Company
Deluxe Package	<25000	31-40	Managers belong to City tier 3/City tier 1 and divorced	Married&Divorced	Small Business	Company
King Package	30000-50000	51-60	VP. Belong to city tier 1	Single	Small Business&Female	N/A
SuperDeluxe	<35000	41-50	AVP, belongs to city tier 3	Single	Salaried Males	Company invited
Standard Deluxe	<30000	31-40	Senior Manager, belongs to city tier 3	Married	Small Business	Self Inquired

The results from Exploratory Data Analysis (EDA) showing the ideal customer profile are young, single professionals (e.g., business owner, managers) who already has a passport and travel history. This indicates the travel experience and assets are strong indicators travel package purchasing analysis. At the same time, customers invited by the company are more likely to buy, especially with a number of follow-ups of 6 to 9, and they will pick a basic package. This illustrates that commercial engagement is also influential. Additionally, customers with a higher city tier are more likely to prefer a luxury package. Conversely, demographic factors such as gender, car ownership, and the number of children have little correlation with our dependent variables as shown in Table 3.

3.1.4 . Split the Dataset/ before Building the Model

The data preparation step involves feature selection, which drops customer interaction features that won't be available for new customers. This is followed by one-hot coding, which converts categorical variables to a numerical format using dummies. Lastly, as the target variable ProdTaken is significantly imbalanced in distribution, it will use stratified sampling by using a stratified parameter to ensure relative class frequencies are approximately preserved in the train and test sets. The training set contains 70% of the dataset, while the testing set contains the 30% remaining.

The algorithms are evaluated using the following metrics: accuracy, recall, precision, and F1-score. These scores quantify the model's predictive accuracy. A confusion matrix is also generated to provide a visualization of the model's performance, particularly its precision and error types.

3.2 . Initial Decision Tree & Random Forest

In building DT algorithm, first it create a decision tree classifier with explicit class weights (20% weight for class 0, and 80% for class 10, and this weighting addresses the class imbalance by giving more importance to the minority group. While the initial random forest creates a Random Forest (RF) classifier with default parameters and sets the random state equal to 1 for reproducibility.

Both methods are overfitting the training data as there are lots of disparities between the train and test set, and their recall scores are not so high. Therefore, the hyperparameter in DT using GridSearchCV to optimize for recall score, it defines the maximum depth of the tree at 7, the minimum samples required at leaf nodes and the maximum leaf nodes to limit the total number. The Grid Search Configuration tests all combinations of the specified hyperparameters and the best-performing model is extracted from Grid Search. At the same time, random forest tuning increases the number of trees to 500, which reduces variance while controlling the maximum depth of the trees, the maximum features, and the number of samples. This aims to reduce the issue of overfitting. In recall score improvement, 5-fold-Cross-Validation is used to provide more reliable performance estimates, which extracts the optimal combination from grid search. As a result, the overfitting in both models has been reduced, it could plot the feature importance plot to indicate features that influence purchasing.

4. Result Analysis

The confusion matrix in the four models reveals a clear trade-off between overfitting and generalization of the model. The initial decision tree and random forest model

achieve a perfect score (100%) on training data but suffer from significant overfitting, leading to unstable and poor performance on unseen test data. While the two tuned models demonstrate a lower total true prediction rate, they have a stronger and more reliable performance on the test set, indicating these two models learn a more generalized pattern from data.

The Table 4 summarizes the overall performance of four constructed models. Based on the evaluation metric on accuracy, recall score, precision and F1 score; the Tuned Random Forest model appears as the most effective model due to the superior balance between training and testing performance. This top-performing model is proven by the highest test F1 score (0.566) and the highest test accuracy (0.812), which indicates the model successfully balanced the trade-off between precision and recall, minimizing the overfitting issue. The Tuned Decision Tree ranks as the second-best performer, with the key strength of a higher test recall (0.663) compared to the Tuned Random Forest model. This means the model is more successful at identifying all relevant positive cases and minimizing false negatives. However, this advantage comes at a cost: it has the lowest precision among the four models on both the training and test sets, leading to a higher number of false positives. This trade-off also lowers the F1 score. This model would be a suitable choice if the company's priority is to capture as many positives as possible, with an acceptable number of false predictions.

Table 4: The Confusion Matrix of 4 Models

	Model 1		Model 2		Model 3		Model 4	
	Initial DT train	test	intial RF train	test	tuned DT train	test	tuned RF train	test
True Negative	2772(100%)	1086(91.2%)	2772(100%)	1165(98.1%)	2245(81.0%)	939(79.0%)	2407(86.8%)	1017(85.6%)
False Negative	0	102(8.8%)	0	23(1.9%)	527(19.0%)	249(21.0%)	365(13.2%)	171(14.4%)
true negative	642(100%)	162(58.7%)	642	134(48.5%)	437(68.1%)	183(66.3%)	409(63.7%)	172(62.3%)
false negative	0	114(41.3%)	0	142(51.5%)	205(31.9%)	93(33.7%)	233(36.2%)	104(37.7%)
Total True Prediction	100%	85.20%	100%	88.70%	78.50%	76.60%	82.40%	81.20%

Table 5: The summary of model performance

Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision	Train_F1	Test_F1
Tuned Decision Tree	0.78559	0.76639	0.68069	0.66304	0.45332	0.42361	0.54421	0.51695
Tuned Random Forest	0.82484	0.81216	0.63707	0.62319	0.52842	0.50146	0.57768	0.55574
Decision Tree	1	0.85246	1	0.58696	1	0.61364	1	0.6
Random Forest	1	0.8873	1	0.48551	1	0.8535	1	0.61894

In contrast, the Standard Decision Tree and Random Forest exhibit a clear sign of overfitting. This large disparity between training and testing metrics is strong evidence of this issue. For example, the low recall score in both models indicates the inability to summarize. Therefore, their high performance on the training data is deceptive and confirms that tuning was essential to build models that perform reliably in the real world as shown in Table 5.

ISSN 2959-6130

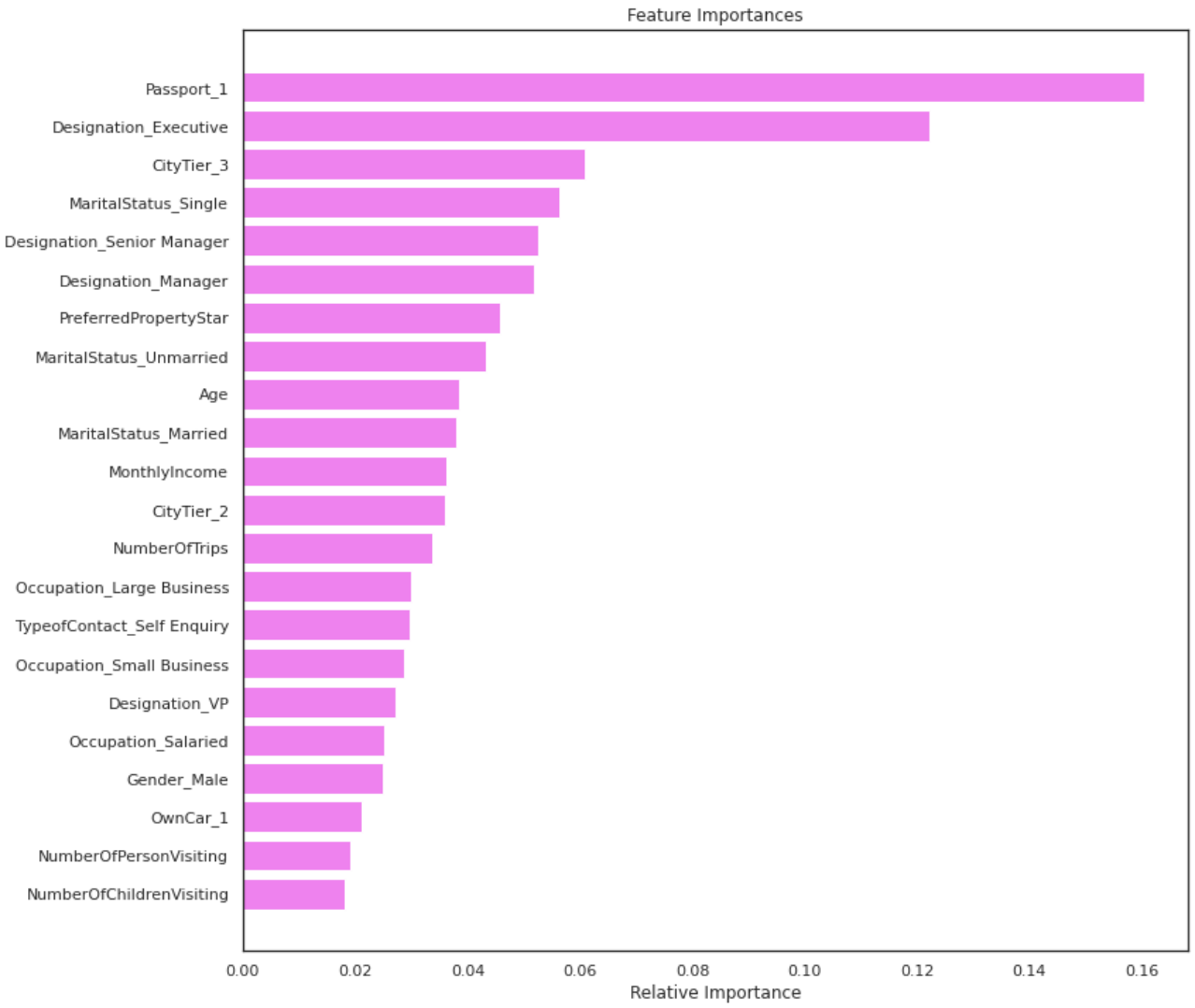

Figure 1: The feature importance plot in tuned DT(Picture credit: Original)

Therefore, it pick Tuned Decision Tree Model and Tuned Random Forest Model for prediction. The most Important features in Tuned DT are passport , Desgination as Executive,City tier 3 from Figure 1 and Figure 2.

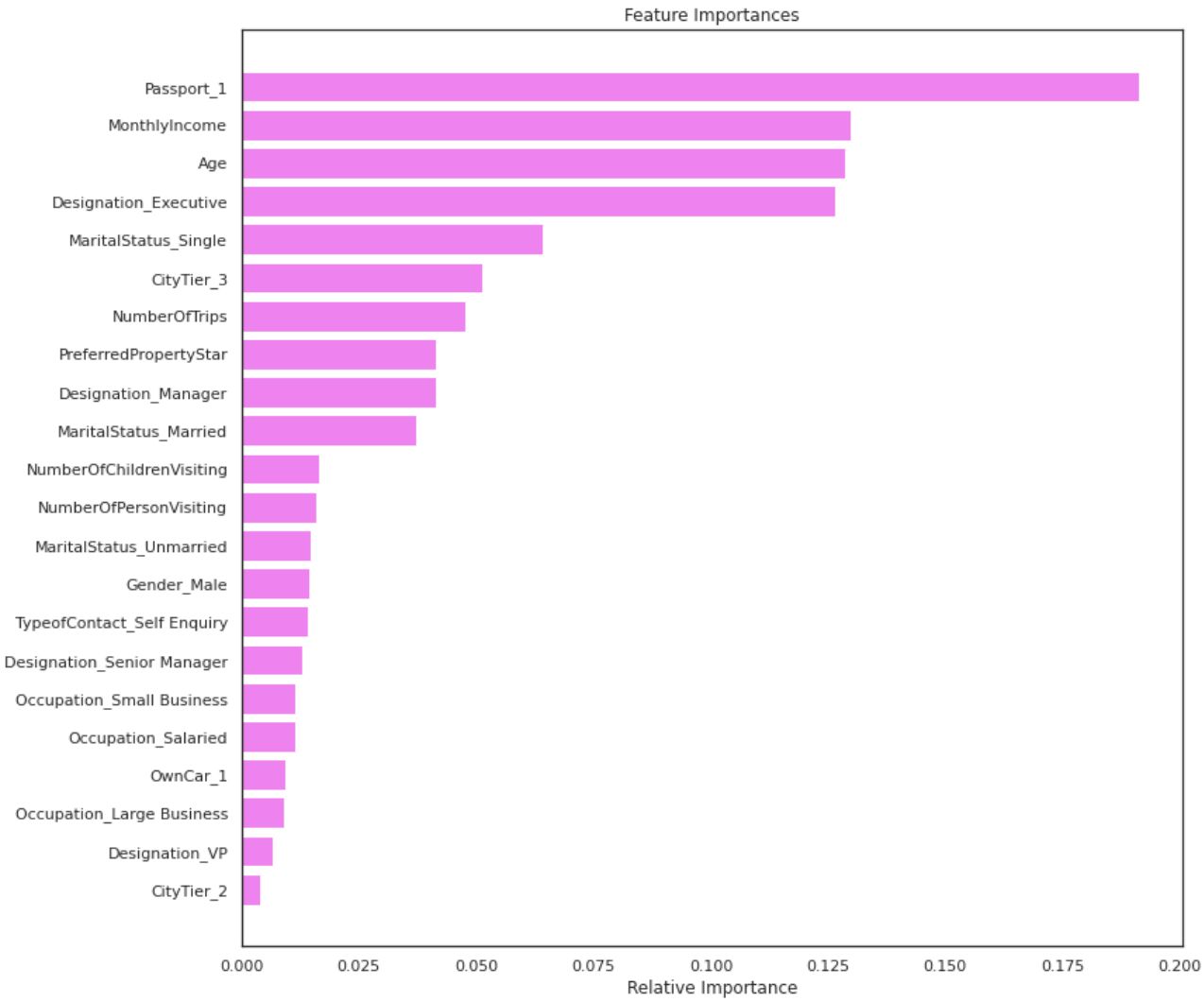

Figure 2: The feature importance plot in tuned RF(Picture credit: Original)

5. Business Implication

Based on the developed predictive model, several strategic business implications could be implemented for optimizing travel package marketing and sales. The models provide valuable insights into key factors driving purchasing decisions of past travel package: with Designation, Passport Ownership, City Tier, Marital Status and Occupation as the most significant predictors. For analyzing these predictors, customers with designation as Executives should be primary target. Within this segment, the highest possibility to purchase for package is observed among customers who hold a passport, in Tier 3 cites and single or unmarried. Additionally, customers with monthly income between 15000 and 25000, has the aged between 15 and 30 and those who prefer 5-star properties also shows a strong likelihood of accepting the new package.

Connecting the results from the EDA and predictive modeling, it found that customer interaction quality plays an important role in conversion rates. Strategies should focus on customers who respond to longer pitch durations, achieve pitch satisfaction scores of 3 or 5, and receive multiple follow-ups, as this segment demonstrates a six-times higher purchase rate. The company should increase spending on targeting this specific segment among both existing and potential customers.

Furthermore, passport ownership consistently appears as a primary feature in all feature importance plots. To leverage this, the company should consider integrating passport or visa application services into new travel packages as a complimentary value-added service. Alternatively, establishing a dedicated department to assist with passport services could be a strategic initiative to increase passport ownership among its customer base, thereby expanding the target market. While single customers dominate current sales, creating family-friendly packages with child-

ISSN 2959-6130

care services could unlock value in the married segment.

6. Conclusion

This paper compares the decision tree and random forest models for predicting customer purchase behavior in the tourism industry. The analysis revealed that while both models required tuning to overcome overfitting, the Tuned Random Forest model demonstrated superior overall performance, achieving the highest balance between precision and recall. These findings confirm established literature on the power of ensemble methods but offer a nuanced business insight: the Tuned RF models with highest prediction power with a balanced level between training and testing set. . This results fits the comparison discussed above which RF is built randomly selected features subspaces, with 100% accuracy on traning data and calculated the average results. In a result, a better performance shown in RF models.

The results provide a data-driven foundation for marketing strategy, identifying the ideal customer as a young, single executive with a passport. For 'Visit with us,' this means reallocating resources to target this demographic and potentially offering integrated passport services. Ultimately, this paper affirms that predictive analytics is a vital asset for strategic decision-making. However, this research is limited to two algorithms and one industry. Future work should broaden this comparison to include models like XGBoost and test these findings in diverse sectors such as finance or e-commerce to further establish their generalizability and business value.

References

[1] Elkan, C. (2013) Predictive Analytics and Data Mining (Vol. 600). San Diego: University of California, 2, 40-46.

[2] Kumar, V. & Garg, M. L. (2020) Predictive Analytics: A Review of Trends and Techniques. International Journal of Computer Applications, 182, 31-37.

[3] Lee, C. S., Cheang, P. Y. S. & Moslehpour, M. (2022) Predictive Analytics in Business Analytics: Decision Tree. Advances in Decision Sciences, 26, 1-29.

[4] Gepp, A., Kumar, K. & Bhattacharya, S. (2010) Business Failure Prediction Using Decision Trees. Journal of Forecasting, 29, 536-555.

[5] Valecha, H., Varma, A., Khare, I., Sachdeva, A. & Goyal, M. (2021) Prediction of Consumer Behaviour Using Random Forest Algorithm. 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering ,2, 1-6.

[6] Hart, P. E., Stork, D. G. & Duda, R. O. (2001) Pattern Classification. Hoboken: Wiley, 32, 140-146.

[7] Shi, H. (2007) Best-First Decision Tree Learning (Doctoral dissertation, The University of Waikato, 23, 45-56.

[8] Ho, T. K. (1995) Random Decision Forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 12, 278-282.

[9] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.

[10] Salman, H. A., Kalakech, A. & Steiti, A. (2024) Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, 69-79.

[11] Kaggle. (2024) Travel Package Prediction: Ensemble Techniques. Available, 16, 75-80.