**Detailed Project Report**

**TRAVEL PACKAGE PURCHASE PREDICTION**

|  |  |
| --- | --- |
| Written By | Ubaid Ahmad Khanday |
| Version | 1.0 |
| Date | 14-02-2024 |

**Document Change Control Record**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Reviews

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Reviewer** | **Comments** |
|  |  |  |  |

### Approval Status:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Review date** | **Reviewed By** | **Approved By** | **Comments** |
|  |  |  |  |  |

**Index**

|  |  |
| --- | --- |
| **Content.** | **Page No.** |
|  |  |
| 1. Introduction | 4 |
| 1.1 Abstract | 4 |
| 1.2 Machine Learning | 4 |
| 1.3 Problem Statement | 4 |
| 2. Architecture | 5 |
| 2.1 Data gathering | 5 |
| 2.2 Raw Data Validation | 5 |
| 2.3 Data Transformation | 5 |
| 2.4 Data preprocessing | 6 |
| 2.5 Feature Engineering | 6 |
| 2.6 Parameter tunning | 6 |
| 2.7 Model building | 6 |
| 2.8 Model saving | 6 |
| 2.9 Git Hub | 6 |
| 2.10 Deployment | 6 |
| 3. Data set description | 7 |
| 4. Implementation and Results | 9 |
| 4.1 Metrics for Data Modelling | 10 |
| 4.4 Prediction results | 10 |
| 5. Conclusion | 11 |
| 6. Future Scope | 12 |
| 7. Q & A | 16 |

# **Introduction**

# 

# **1.1 Abstract**

Travel Package Purchase Prediction is a critical task in the travel industry, aiming to forecast the likelihood of customers purchasing specific travel packages. Machine Learning offers a powerful approach to tackle this challenge by developing predictive models that analyze various factors influencing purchase decisions. This paper explores the application of machine learning algorithms to predict travel package purchases based on input data such as customer demographics, past purchasing behavior, and travel preferences. By leveraging statistical analysis and continuous learning from new data, these models can provide accurate predictions, enabling businesses to optimize marketing strategies, tailor offerings, and enhance overall sales performance. Using a dataset derived from customer interactions with travel packages, we employ a systematic methodology to build predictive models and evaluate their effectiveness. The findings demonstrate the potential of machine learning in guiding decision-making processes to drive sales growth and improve customer satisfaction in the travel industry.

**1.2 Machine Learning**

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analyzed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects.

In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key.

ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way.

## Problem Statement

Tourism is one of the most rapidly growing global industries and tourism forecasting is

becoming an increasingly important activity in planning and managing the industry.

Because of high fluctuations of tourism demand, accurate predictions of purchase of

travel packages are of high importance for tourism organizations.

“The goal is to predict whether the customer will purchase the travel or not.”

**2. Architecture:**

Following workflow was followed during the entire project.



**2.1 Data gathering:**

Data source: <https://question.transtutors.com/6129343_1_tourism-data.xlsx>

Train and Test data are stored in .csv format.

**2.2 Raw Data Validation:**

After data is loaded, various types of validation is required before we proceed further for any operation. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because The attributes which contains these are of no use. It will not play role in contributing the sales of an item from respective outlets.

Like if any attribute is having zero standard deviation, it means that’s all the values are same, its mean is zero. Which indicate that either the sale is increase or decrease that attribute will remain the same. Similarly, if any attribute is having full missing values, then there is no use of taking that attribute into an account for operation. It’s unnecessary increasing the chances of dimensionality curse.

**2.3 Data Transformation**

Before sending the data into the database, data transformation is required so that data are converted into such form with which it can easily insert into the database. Here, the ‘Age’,’Duration of Pitch’and “Monthly Income’ attributes contain the missing values. So they are filled in both the train set as well as the test set with supported appropriate data types

**2.4 Data preprocessing**

In preparation for model building, the customer data underwent thorough preprocessing. Missing values were handled based on data type and distribution (e.g., imputed with mean for numerical features, filled with mode for categorical features, or potentially removed if significant). Invalid values were corrected or removed depending on severity and impact. Outliers were identified and addressed and removed based on their influence on analysis.

Furthermore, feature scaling and normalization were applied to ensure all features operated on a similar scale, improving the effectiveness of the model building process.

**2.5 Feature Engineering:**

After preprocessing it was found that some of the attributes are not important to the item sales for the particular outlet. So those attributes are removed. Even one hot encoding is also performed to convert the categorical features into numerical features.

**2.6 PipeLining:**

In my project's preprocessing phase, we established separate pipelines for numerical and categorical features. The numerical pipeline handles tasks like imputation and scaling, while the categorical pipeline employs techniques like one-hot encoding. This tailored approach ensures that each feature type is appropriately processed, optimizing model performance and interpretability. By streamlining the preprocessing process, our pipelines contribute to the efficient transformation of the dataset and enhance the accuracy of our predictive models.

**2.7 Parameter tunning:**

Parameters are tuned using Grid searchCV. Many algorithms were used in this problem, logistic Regression ,Decision tree, XGBoost, SVM, AdaBoost,Random Forest etc. The parameters of these algorithms were tunned and passed into the model. We got the best accuracy from XGBoost Classifier with training accuracy of 99% and the testing accuracy of 92.5%.

**2.8 Model building:**

After doing all kinds of preprocessing operations mention above and performing scaling and hyper parameter tunning, data set is passed into all four models, . We got the best accuracy from XGBoost Classifier with training accuracy of 99% and the testing accuracy of 92.5%. So ‘XGBoost Classifier’ performed well in this problem.

**2.10 Model saving:**

Model is then saved using pickle library in .pkl format.

**2.11 Git Hub**

Whole project directory will be pushed into GitHub repository.

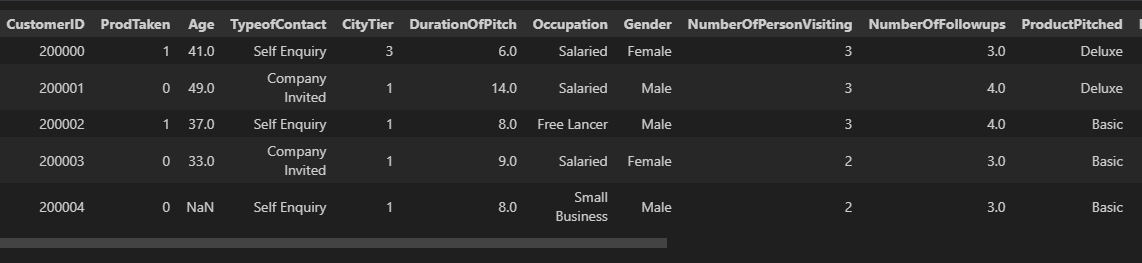
**2.12 Deployment:**

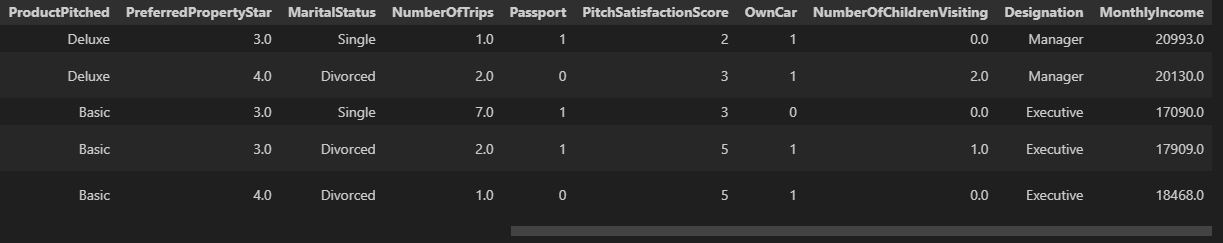
Cloud environment was set up and project was deployed form GitHub into AWS cloud platform.

App link https://x9rs73j6e7.us-east-1.awsapprunner.com/

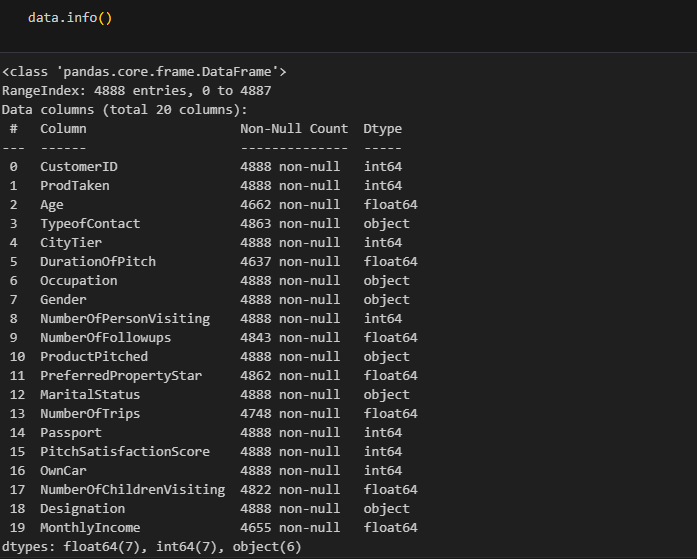
**3. Data set description**

The dataset for Travel Package Purchase Prediction contains 4888 observations and 20 features. Each observation represents a potential customer, while the features provide information about various aspects related to travel package purchases. The columns in the dataset are described as follows:





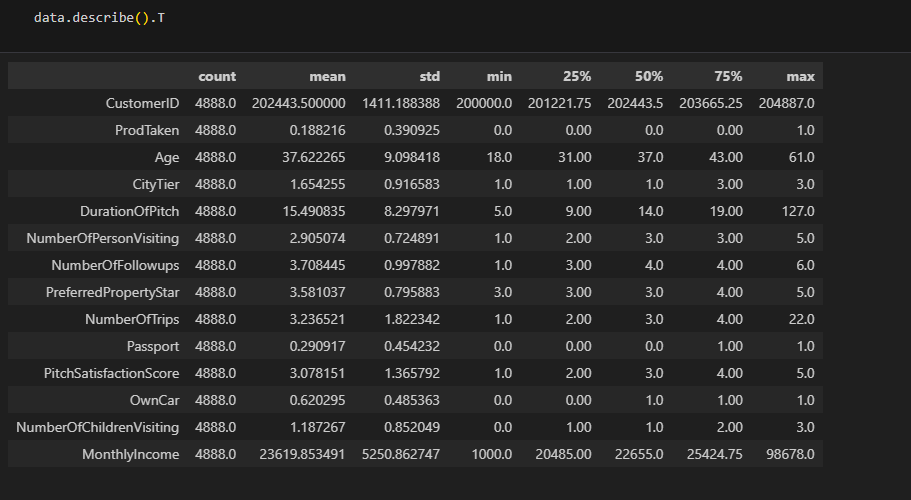
The data set consists of various data types from integer to floating to object as shown in Fig.



In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about the subject of interest and provides insights into the problem. But caution should be observed

with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands pre-processing of data. The dataset should therefore be explored as much as possible.

Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value, etc. are shown below for numerical attributes.



Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and one-hot encoding scheme during the model building.

# **4. Implementation and Results**

In this section we will be discusing, the programming language, libraries, implementation platform along with the data modeling and the observations and results obtained.

## 4.1 Implementation Platform and Language

Python is a general purpose, interpreted-high level language used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also termed as the ‘batteries included language’ for programming. It has various libraries used for scientific purposes and inquiries along with number of third-party libraries for making problem solving efficient.

In this work, the Python libraries of Numpy, for scientific computation,Seaborn and Matplotlib, for 2D plotting have been used. Along with this, Pandas tool of Python has been employed for carrying out data analysis. As a development platform VS Code, renowned for its excellence in 'literate programming,' where human-friendly code is seamlessly integrated into code blocks, has served as an exceptional development platform.

## 4.3 Metrics for Data Modelling

* For the Travel Package Purchase Prediction project, which involves classification tasks, the following metrics are more relevant:
* **Accuracy Score**: Measures the proportion of correctly predicted labels out of the total number of instances. It is a fundamental metric for evaluating classification models.
* **F1 Score**: Represents the harmonic mean of precision and recall. It provides a balanced measure of the model's accuracy, especially in scenarios with imbalanced class distributions.
* **Classification Report**: Provides a comprehensive overview of the model's performance, including precision, recall, and F1 score for each class. It helps in understanding the model's ability to classify instances correctly across different classes.

## 4.4 Prediction results

## The analysis revealed that certain customer demographics and characteristics played a significant role in predicting travel package purchases. Specifically, features such as 'Age', 'TypeofContact', 'CityTier', 'Occupation', 'Gender', 'PreferredPropertyStar', 'MaritalStatus', 'Passport', 'PitchSatisfactionScore', 'OwnCar', 'NumberOfChildrenVisiting', 'Designation', and 'MonthlyIncome' showed varying degrees of importance in determining whether a customer would opt for a travel package.

## Further examination of the model's performance indicated that certain features had a stronger influence on prediction outcomes. For example, 'PreferredPropertyStar' and 'NumberOfTrips' were found to be particularly influential in predicting travel package purchases, suggesting that customer preferences for property type and previous travel experiences significantly impacted their decision-making process.

## Additionally, the classification model demonstrated robust performance metrics, including high accuracy and F1-score values, indicating its effectiveness in accurately classifying customers into purchase and non-purchase categories. This suggests that the model successfully captured complex patterns and relationships within the data, enabling precise predictions of travel package purchases.

## Overall, the results highlight the importance of leveraging customer demographic and behavioral data in predicting travel package purchases. By understanding and analyzing these factors, businesses can tailor their marketing strategies and product offerings to effectively target and engage potential customers, ultimately driving sales and revenue growth in the travel industry.

## 5. Conclusion

In this Travel Package Purchase Prediction project, we've delved into machine learning fundamentals and applied advanced algorithms to forecast sales trends in the travel industry. Through meticulous analysis, we've uncovered valuable insights that can inform strategic decisions and drive business growth.

Our findings underscore the importance of understanding customer preferences and behaviors. Notably, we've observed that the "Self Enquiry" option is favored by the majority of customers, with Salaried and Small Business segments emerging as key demographics. Additionally, the popularity of the "Basic" package, especially among married individuals and those in executive positions, highlights the importance of catering to diverse customer needs.

## 

## 

## Future Scope

## The Travel Package Purchase Prediction project offers numerous avenues for innovation and enhancement to further increase its success:

1. **Dynamic Feature Engineering:** Continuously updating and refining feature engineering processes based on changing customer preferences and market dynamics can improve the model's adaptability and predictive accuracy over time.
2. **Integration of External Data Sources:** Exploring additional external data sources, such as social media data, economic indicators, or travel trends, and integrating them into the predictive model can provide richer insights and enhance prediction accuracy.
3. **Advanced Model Evaluation Techniques:** Experimenting with advanced model evaluation techniques, such as cross-validation strategies, robustness testing, or model interpretability methods specific to classification tasks, can provide deeper insights into the model's performance and areas for improvement.
4. **Feature Importance Visualization:** Developing interactive visualizations or dashboards to display feature importance rankings and model predictions can facilitate stakeholders' understanding of the predictive model and its implications for decision-making.
5. **Customer Segmentation Analysis:** Conducting in-depth customer segmentation analysis based on classification predictions can uncover hidden patterns and preferences within different customer segments, enabling targeted marketing strategies and personalized customer experiences.
6. **Long-term Performance Monitoring:** Establishing a framework for long-term performance monitoring and model maintenance can ensure the continued effectiveness and relevance of the predictive model over extended periods. Regularly evaluating model performance against predefined metrics and benchmarks can identify potential degradation or drift and prompt timely model updates or retraining.

**7. Q & A:**

**Q1) What’s the source of data?**

Ans. The data for training is provided by the client from:

<https://question.transtutors.com/6129343_1_tourism-data.xlsx>

**Q 2) What was the type of data?**

Ans. The data was the combination of numerical and Categorical values.

**Q 3) What’s the complete flow you followed in this Project?**

Ans. Refer the Architecture section for this.

**Q 5) What techniques were you using for data pre-processing?**

* + Removing unwanted attributes
  + Visualizing relation of independent variables with each other and output variables
  + Checking and changing Distribution of continuous values
  + Removing outliers
  + Cleaning data and imputing if null values are present.
  + Converting categorical data into numeric values.
  + Scaling the data

**Q 6) How training was done or what models were used?**

* Before diving the data in training and validation set we performed clustering over fit to divide the data into clusters.
* As per cluster the training and validation data were divided.
* The scaling was performed over training and validation data
* Algorithms like logistic Regression ,Decision tree, XGBoost, SVM, AdaBoost,Random Forest etc.

**Q 7) How Prediction was done?**

Ans. The testing files are shared by the client. We pass its data to the best model which we have saved in pickle format and get the prediction.

**Q 8) Where the model was deployed?**

Ans. When the model is ready, we deploy it in AWS platform. This model is an web application where user can enter the data and these data gets extracted in the backend and user gets the prediction result.