# SAS: Salesperson Assisting System

# INTRODUCTION

Machine learning (ML) is the application of artificial intelligence for a computer to perform human tasks on its own (IBM, 2020). ML is applied to many industries and one of them is tourism. In this project, we will apply ML to a tourism company to alleviate the problem of high marketing costs and a low package sign-up rate. We will use ML to segment the customer for targeted advertising and recommend tour packages to alleviate the problem. Given that the tourism sector is among the most affected by the COVID-19 pandemic, the application of ML here will not only solve this problem but also accelerate the recovery of the company from the pandemic.

## Problem Statement

Visit with Us is an Indian company in the wellness tourism market. The main problem with the company is the high marketing cost with a low package sign-up rate. According to the existing data, 30% of customers have low satisfaction with the pitched package whereas only 18% of customers bought the recommended package*.*

The sub-problem that contributes to the main problem is 1) there is no targeted advertising because customers are contacted at random 2) Once a customer is contacted, the tour packages were chosen at random to be pitched to the customer. Therefore, to address the main problem of high marketing costs with a low package sign-up rate, we will need to address the sub-problem. Addressing the sub-problem forms a basis for our objective.

## Objective

To address the problem, the objectives are:

1. To determine significant features for package selection using the feature selection algorithm.
2. To identify customer segment with a higher sign-up probability for advertising using clustering algorithm.
3. To develop a package recommender to suggest tour packages to be pitched using collaborative filtering.
4. To evaluate the package recommender’s performance.

## Question

To address the objectives, the questions are:

1. What are the significant features for package selection using the feature selection algorithm?
2. Which customer segment will have a higher probability of signing up?
3. How to develop a package recommendation based on significant features identified?
4. Which package recommender is better?

## Deliverable

Once we address the question, the deliverables are:

1. Significant features for package selection.
2. Customers segments with a higher probability of signing up.
3. A set of package recommenders based on significant features.
4. The optimal package recommenders to be used.

## Motivation

The tourism sector is affected by the COVID-19 pandemic. The pandemic measures implemented across the world have reduced international arrivals by 70%. 100 million direct tourism jobs are at risk, and small businesses are particularly vulnerable (World Tourism Organization, 2022). As the countries are relaxing the travel restriction and opening the border, the tourism sector is slowly resuming normal operations. The tour operator needs to have a robust strategy to accelerate their recovery to remain financially viable. This study will help the tour operator to identify their target customer segment and turn them into a customer by pitching the most suitable package based on experience.

## Significance

This study provides customer segments that the marketing team can focus on in their marketing strategy. Targeted marketing can help the business to identify a specific group of people who are interested in their tour package and attract high-quality leads that can turn into potential customers. Besides, by delivering a recommender system that aids the salesperson in selecting the tour package to pitch instead of choosing randomly, it will also improve the tour package taking up rate and help the recovery of the tourism industry.

# BACKGROUND/LITERATURE REVIEW

## Feature Selection

Feature selection is a data pre-processing strategy to select a significant subset of features from the original dataset by identifying and eliminating redundant and irrelevant features. There are three types of features selection methods: 1. Filter method. 2. Wrapper method 3. Embedded method.

The filter method selects features irrespective of the nature of the classifier employed. This method has the benefit of being simple and independent of the type of classifier employed. However, this technique has the drawback of ignoring interactions with the classifiers and feature dependencies.

The wrapper method is reliant on the classifier employed, i.e., the classifier's output is used to assess the quality of the provided feature or attribute. This method has the benefit of eliminating the drawbacks associated with the filter method. However, the downside of this method is that it is comparatively slower than the filter method due to the processing time of feature dependencies. The classifier's performance directly measures the feature's quality.

The embedded method seeks the best subset of features that are built into the classifier. This method has the benefit of being less computationally demanding than the wrapper method (Sekhar & Sujatha, 2020).

## Market Targeting

Market targeting is the process of determining potential customers, selecting which group of customers to pursue, and creating value for the targeted customers. For this project, the process can be broken down into two steps: segmentation and targeting (Corporate Finance Institute, 2022).

Segmentation is about grouping customers with similar needs together and then finding out the characteristics and factors they share (D’Urso et al., 2021). The different type of segmentations are geographic, demographic, psychographic, and behavioral variable (APELO Consulting, 2022). There are many analysis methods used for segmentation: k-means clustering, two-step cluster analysis, latent class analysis, and latent class choice modelling (APELO Consulting, 2022).

After the segmentation is completed, we will perform market targeting. A market target is the process of evaluating the attractiveness of the segment. In this step, the company considers the characteristics of the segment such as its profitability of the segment (Corporate Finance Institute, 2022).

## Recommendation System

### Overview

A recommendation system is a system that selects a subset of items that is relevant to a user given set of items. A recommendation system has four major types: collaborative filtering system, content-based system, hybrid system, and knowledge-based system. In a collaborative filtering system, the system uses responses from multiple users to make a recommendation. In a content-based system, the system uses items and user attributes and responses from a single user to make a recommendation. A hybrid system is a system that combines collaborative filtering and a content-based system. In a knowledge-based system, the user specifies the criteria, and the system recommends according to the criteria (Aggarwal, 2016). In this literature review, we focus only on the collaborative filtering system as it is the only system among the four that is relevant to our problem.

A collaborative filtering system has two types: neighbour-based system and model-based system. A neighbour-based system is a system that makes a recommendation for a user based on the response from another user on an item. A model-based system is a system that creates a model based on the sparsely populated m x n matrix and makes recommendations using the model (Aggarwal, 2016).

### Neighbour-based Collaborative Filtering

***Overview.*** The neighbour-based collaborative filtering system is a generalized k-nearest neighbour (kNN) classifier. The procedure follows a kNN classifier: 1. Compute the similarity of every user relative to the input user using a *similarity function*. 2. Get the top k similar user to the input user 3. For each item, predict a rating for the given user based on the rating from the top k most similar user using a *prediction function.* This is the rating predicted by the system for the input user on an item. 4. Get items with top m ratings. These are items recommended by the system to the given user (Aggarwal, 2016).

***Similarity Function*.** The accuracy of the system depends on whether the similarity function finds users that are similar enough to the given users. Traditionally, the similarity function used is either cosine similarity or Pearson’s correlation (Aggarwal, 2016). Fuzzy similarity can also be used as the similarity function to improve the accuracy of the similarity function (Houshmand‐Nanehkaran et al., 2022).

***Prediction Function*.** The accuracy of the system depends on how the prediction function takes in the rating of a similar user and outputs a predicted rating for the user on an item. Traditionally, the prediction function is the average of the rating weighted by the similarity of the user (Aggarwal, 2016). Compared to the similarity function, there are fewer attempts to improve the prediction function. A TOPSIS solution to improve upon the traditional method (Al-bashiri et al., 2018).

***Optimization***. It is known that if we use the procedure above directly, the computation complexity is high. One method to mitigate the issues is to use clustering. Specifically, during the training phase, we can use a clustering algorithm to assign each row in the training data to a cluster. During the testing phase, we can find out which cluster does the input row belongs to and compute the similarity between the input row and the row in the training dataset in the cluster instead of computing the similarity between the input row to every single row in the training dataset (Aggarwal, 2016). Another method to mitigate the issues is to use KD trees. This method uses trees to encode the distance information in the sample such that if we know A and B are far apart and B and C is close together, then A and C must also be far apart. The use of this data structure helps to cut down redundant computation when finding the nearest neighbour (scikit-learn, 2022).

### Model-based Collaborative Filtering

***Overview.*** The model-based collaborative system is an approach to recommender system where a model of data is developed. Model-based collaborative filtering can be developed using two approaches using probability or rating prediction (Aditya et al., 2016). For instance, the Bayesian classifier is an example of a probability approach and Singular Value Decomposition, Markov Decision Process and latent semantics are a type of rating prediction.

***Matrix Factorization.*** The matrix factorization (MF) algorithm is a latent factor model that relies on the fact that any rating matrix is a multiplication of two-component matrices: a user x latent factor matrix and a latent factor x item matrix (Funk, 2006). These two-component matrices need to be estimated from the data so that the multiplication of the component matrix is as close to the observed matrix as possible. Once the optimal component matrices are obtained, the prediction of the rating from user *i* on item *j* is the dot product of row *i* in user x latent factor matrix and column *j* in latent factor x item matrix. SVD++ improve on MF by including implicit information (Koren, 2008). A new factorization method called Bernoulli matrix factorization is also being proposed (Ortega et al., 2021).

***Decision Tree.*** Aggarwal (2016) proposed using a decision tree as model-based collaborative filtering. In this approach, each item will have its decision tree that predicts its rating and other items will be used as the independent feature. To deal with the sparse matrix issue, Aggarwal (2016) suggests using PCA on the matrix to mitigate the issue.

## Recommendation System Evaluation

Offline evaluations are the most common evaluation method for a recommender system. The purpose of the offline evaluation is to select the best model to deploy online (Gebremeskel & de Vries, 2016). In classification models like logistic regression and K-nearest neighbours, in addition to accuracy, precision and recall are the common metrics used (Al-bashiri et al., 2018; Ortega et al., 2021). F-measure and coverage are also being used in evaluating the performance of different collaborative filtering models (Kim et al., 2019). However, it is worth taking note that offline evaluation system relies on historical data can result in bias. Such bias is often linked to the data collection process where the user interaction information is Missing Not at Random (MNAR) (Carraro & Bridge, 2022).

# METHODOLOGY

## Overview

Our dataset is from Kaggle.com (Kaggle, 2022). The dataset has 17 columns and 4888 rows. The table below shows the metadata. The procedure is 1) pre-process the data. 2) select features that is related to ProdTaken. 3) use these features to i) identify customer segments with a higher probability of ProdTaken, ii) develop and evaluate the recommendation system.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dimension | Column Name | Data Type | Description | Categories |
| Identifier | Customer\_ID | Nominal | Identifier |  |
| Demographic | Age | Ratio | Age |  |
|  | Occupation | Nominal | Occupation | Small Business, Large Business, Free Lancer, Salaried |
|  | Gender | Nominal | Gender | Female, Male |
|  | MaritalStatus |  | Marital status | Single, Unmarried, Divorced, Married |
|  | MonthlyIncome | Ratio | Monthly income |  |
| Travel Behaviour | NumberOfPersonVisiting | Ratio | Number of persons travel with the customer |  |
|  | NumberOfChildrenVisiting | Ratio | Number of children (age < 5) travel with the customer |  |
|  | PreferredPropertyStar | Ordinal | Preferred hotel property rating | 1,2,3,4,5 |
|  | NumberOfTrips | Ratio | Average number of trips per year |  |
| Other details | Passport | Nominal | Customer has a passport or not | No (0), Yes (1) |
|  | Designation | Ordinal | Designation of the customer at work | Executive, Manager, Senior Manager, AVP, VP |
|  | CityTier | Ordinal | Tier depends on the city’s development, population, facilities, living standards | Tier 1(1), Tier 2(2), Tier 3(3) |
|  | OwnCar | Nominal | Customer has a car or not | No (0), Yes (1) |
| Customer interaction | TypeofContact | Nominal | How customer was contacted | Company Invited or Self Inquiry |
|  | NumberOfFollowups | Ratio | Number of follow-ups by the salesperson after the pitch |  |
|  | DurationOfPitch | Ratio | Duration of the pitch by a salesperson |  |
|  | PitchSatisfactionScore | Ratio | Sales pitch satisfaction score |  |
| Target | ProdTaken | Nominal | Customer has purchased a package or not | No (0), Yes (1) |
|  | ProductPitched | Nominal | Product pitched by the salesperson | Basic, Standard, Deluxe, Super Deluxe, King |

**Table 3.1 Metadata**

## Data Pre-processing

Data pre-processing remove incomplete and inconsistent raw data for better analysis result. The methods for data pre-processing are: 1) Import the necessary libraries and dataset. 2) Check and handle missing values by dropping or imputing the values with the central tendency. 3) Handle categorical values by recoding them into an integer. 4) Scaling the variables. 5) Split the dataset into a training set and test set in a ratio of 80:20.

## Feature Selection

We use the filter-based feature selection to remove the redundant irrelevant features relative to the ProdTaken by the customers. The selected features from the dataset will be used for the later steps in the procedure.

## Market Targeting

The procedure for customer segmentation algorithm for market targeting are as below: 1) Study and comprehend attributes in the dataset without neglecting knowledge from business domain. 2) Convert categorical attributes into numeric data type 3) Perform correlation analysis on top of attributes available 4) Normalize the dataset by scaling it, removing highly correlated attributes, and bin DurationofPitch. 5) Find optimal number of clusters, k. 6) Perform k-mean clustering 7) Describe our segmented customers.

## Recommendation System

### Overview

We use collaborative filtering because 1) each customer is pitched with one product. Content-based and hybrid filtering is not possible. 2) there is no information on the product. Knowledge-based filtering is not possible. We use ProdTaken and ProductPitched to form five binary target variable that indicate whether the user bought a specific product. The recommendation model will predict these target variables and recommend a product based on the prediction. We oversampled the training dataset using SMOTE so that the product target variables are balanced. We build both neighbour-based and model-based collaborative filtering to see which perform better.

### ***Neighbour-based Collaborative Filtering***

To build and train the model, we use a generalized kNN classifier by Aggarwal (2016). We store the training dataset for prediction. Unlike Aggarwal (2016), we also build and store KD trees with Euclidean distance as the similarity function to speed up the prediction (scikit-learn, 2022).

To make the model predict the target variable, we use the standard procedure provided by Aggarwal (2016) with KD trees. Specifically, for each row of the test dataset, for each product, we compute the similarity and get *k* data point in the training dataset most like the row using KD trees. Then, we get the vote from the similar user on the target variable weighted by similarity. The predicted class for a target variable is the class with the most votes. To make the model recommend, the model selects the product with the highest vote and recommends it to the user.

### Model-based Collaborative Filtering

To build and train the model, we decided to adapt Aggarwal (2016)’s decision tree and regression method. We train a Logistic Regression (LR) classifier for each target variable where features other than the target variable is the independent variable. Unlike Aggarwal (2016), we do not use the decision tree as the classifier because we need to know which package has the maximum probability to be bought by the customer. We also do not use PCA because our data is not sparse.

To make the model predict the product target variable, we inputted the independent variable into the five classifiers and each classifier predict the probability of buying each product. The predicted class for the binary target variable is the class with highest probability. To make the model recommend, the system select the item with the highest probability of buying the item and recommend it to the user.

## Recommendation System Evaluation

We use the predicted and actual product target variable to construct the confusion matrix and compute the accuracy, precision, recall, f1-measure to evaluate the performance of the model. Since there are five target variables, we will have five set of confusion matrix and the standard measures. Because the testing dataset are imbalances, and both Type I and Type II error are equally important, we use macro-average. Finally, the standard measures of each of the tour packages will be added up using weighted average to evaluate the performance of the model. Comparison will then take place to select the best model for recommendation system.

# RESULTS

## Descriptive Statistics

From the data summary, it was identified that the data contained age range between 18 to 61 years old with averages of age being 37, duration of pitch as 15 minutes, number of trips 3 times and monthly income of 23,619. In total, 920 had signed up for tour packages with the majority contacted via self-inquiry.

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

**Figure 4.1 The Shape and Spread for Numerical Data Distribution**

The spread of age shows a normal distribution without any outliers in Figure 4.1. Apart from that, outliers were observed in other variables such as DurationOfPitch (greater than 120), NumberOfPersonVisiting (5) and average NumberOfFollowups (6). Variables that were skewed to the right include NumberOfTrips and MonthlyIncome.

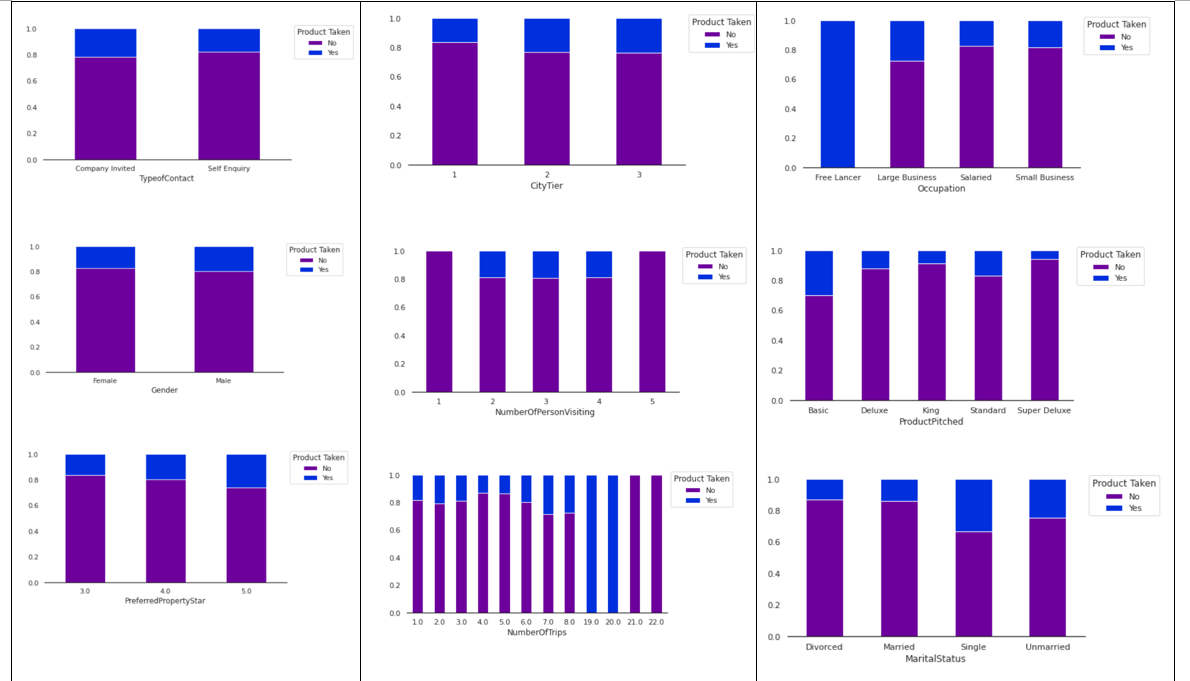
Chart

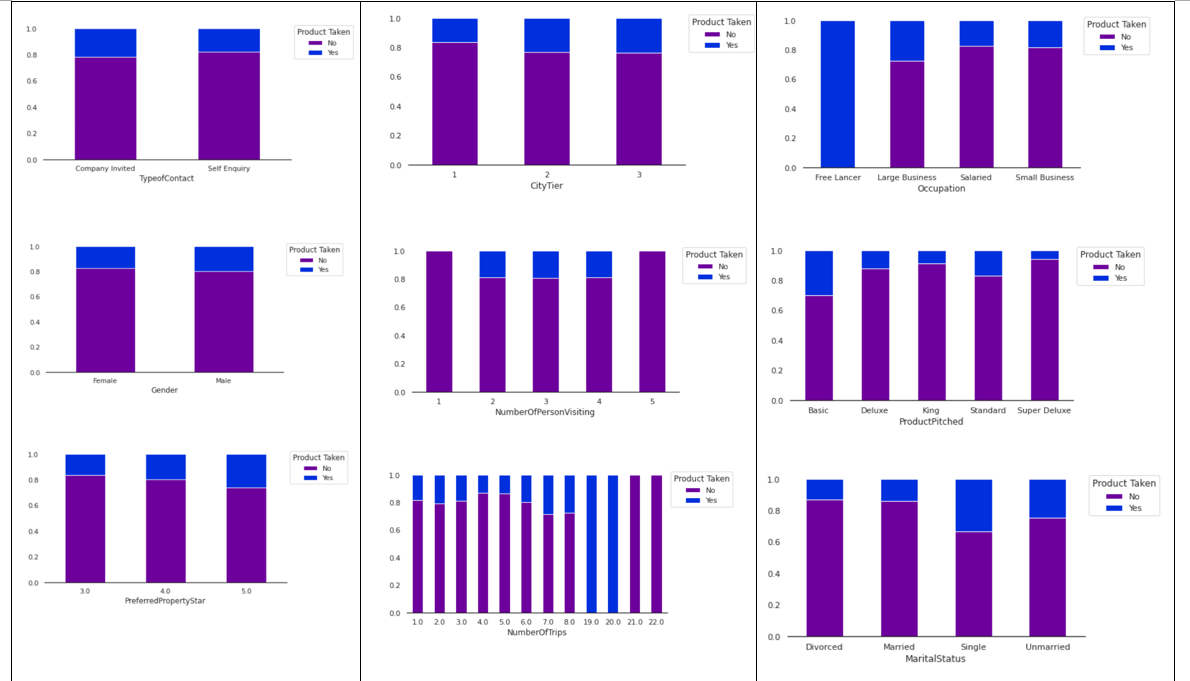
Description automatically generatedChart

Description automatically generated

**Figure 4.2 Nominal Variables**

From Figure 4.2 it is observed that age group and income range were converted into categorical variables as they are vital factors in the tour package selection. Most of the customers who purchased this package fall in the 25-50 age range with monthly salary of 15000-30000. We can also observe that 38% of customers are Executive whereas 35% are managers purchasing this tour package.

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

**Figure 4.3 Stacked Bar Charts**

According to Figure 4.3, variables such as Gender and OwnCar are insignificant to the product taken by the customers. On the other hand, the NumberofTrips between 7 to 8 and 19 to 20 times are more likely to purchase tour packages. Similarly, individuals that owned passport have a higher likelihood of purchasing the product.

## Feature Selection

After preforming data pre-processing and EDA, the features such as CustomerID, Gender, NumberOfChildrenvisiting, and OwnCar are dropped as they are seemed to be insignificant to the model’s performance.

## Market Targeting

K-mean clustering algorithm is sensitive to outliers, thus binned attributes will be the utilized. PCA is not implemented because it will not help us to communicate our findings, even though it can assist us to reduce the number of features in dataset for improved result. Correlation analysis is performed to select attributes and those below 0.7 removal benchmark were excluded namely, Age, MontlyIncome, Gender, Designation, NumberOfChildrenVisiting, OwnCar and DurationOfPitch. The features were categorised into three main segments, demographic, behavioural and psychographic and relevant attributes were assigned to each segment. To determine the optimal number of clusters in each segment, the elbow method was used which uses the total within-cluster sum of square values. As a result, the optimal clusters were assigned as five, five and four for demographic, behavioural and psychographic segments respectively.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

**Figure 4.4 Optimal Number of K-clusters by Pre-determined Segments**

The identified market segments are further explored in Tables 4.1, 4.2 and 4.3 below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Segments | 1 | 2 | 3 | 4 | 5 |
| TypeofContact | Self-Enquiry | Self-Enquiry | Self-Enquiry | Self-Enquiry | CompanyInvited |
| No. of Follow-ups | 4 | 3 | 4 | 4 | 3 |
| Product Pitched | Basic | Deluxe | Basic | Deluxe | Deluxe |
| No. of Trips | 3 | 3 | 2 | 2 | 3 |
| Pitch Satisfaction Score | 4 | 3 | 3 | 3 | 5 |
| DurationofPitchbin | 5-9 | 5-9 | 10-14 | 10-14 | 10-14 |
| ProdTaken | 0 | 0 | 0 | 1 | 0 |

**Table 4.1 Identified Clusters for Behavioural Segment**

In the **behavioral segments** have had 3 to 4 number of trips with Visit with Us, with minimum pitch satisfaction score of 3, duration of pitch that ranges from 5 to 14 minutes and 3 to 4 times number of follow up attempts. It can be interpreted that clusters one and two are urgent and price sensitive potential buyers, clusters three and four are bargain hunters and cluster 5 is status seeking potential buyers. Existing customers only make up one cluster, four. To increase potential sales, Visit with Us will need to conduct market research to understand competitor offering to establish better market positioning to attract clusters with the highest frequency, one and two. Furthermore, the company should focus on cluster 5 as company trip may provide higher profit due to high visitors' head counts. To secure recurring corporate clients, Visit with Us can become a panel tour package provider that will be extremely profitable for the company.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Segments | 1 | 2 | 3 | 4 | 5 |
| Occupation | Small Business | Salaried | Salaried | Small Business | Small Business |
| Marital Status | Unmarried | Married | Single | Unmarried | Married |
| Agebin | 41-50 | 31-40 | 26-30 | 26-30 | 31-40 |
| Incomebin | 20000-25000 | 20000-25000 | 20000-25000 | 20000-25000 | 20000-25000 |

**Table 4.2 Identified Clusters for Demographic Segment**

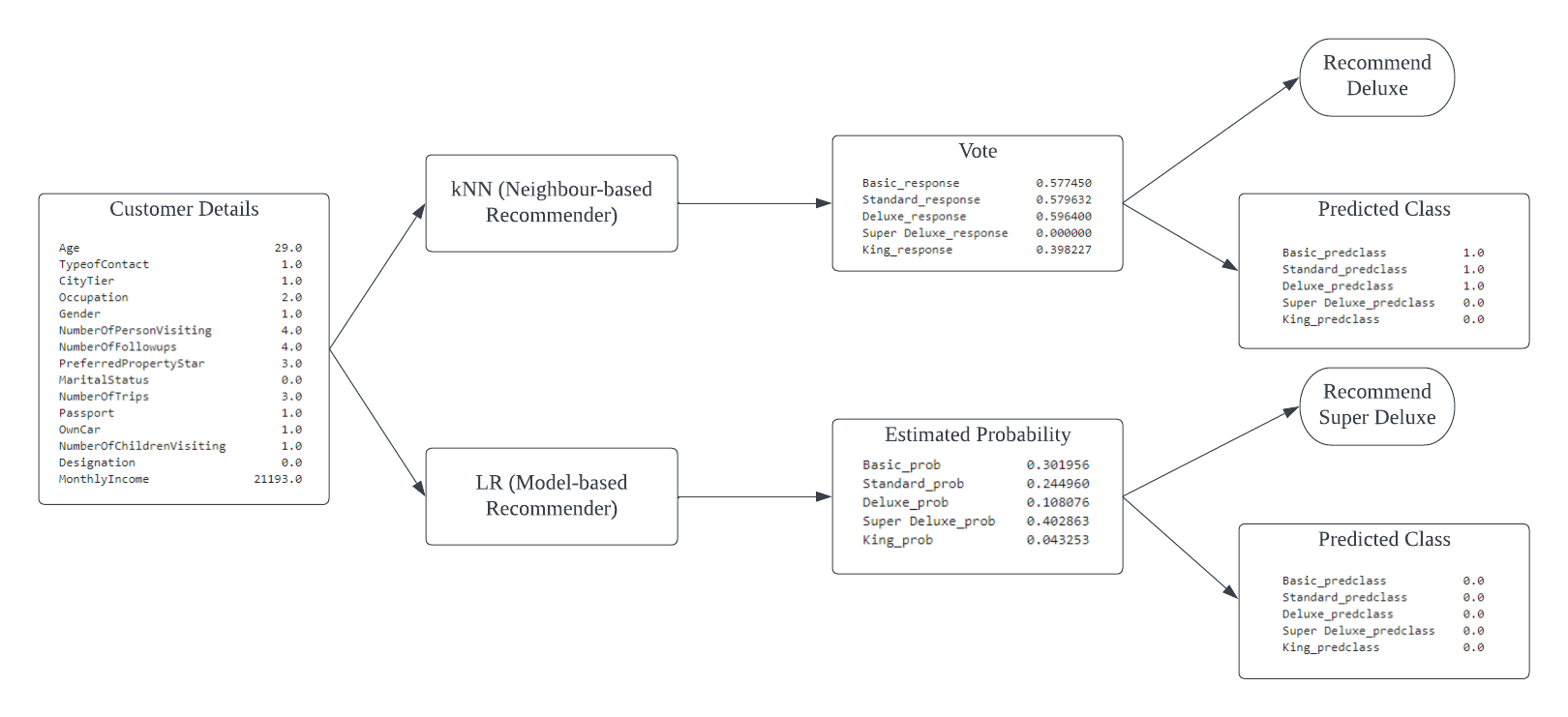
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segments | 1 | 2 | 3 | 4 |
| City Tier | 3 | 1 | 1 | 1 |
| PreferredPropertyStar | 3 | 5 | 3 | 4 |

**Table 4.3 Identified Clusters for Psychographic Segment**

Based on the identified **demographic segments**, the company may focus on expanding its business to cluster 4 as they have potential to be lucrative customers. In view that clusters 1 and 3 of the **psychometric segment** have the highest count and share low budget attributes, Visit with Us should focus to use its budget-friendly package as a good market penetration strategy to their potential customers. Once these potential customers have become their existing customers, Visit with Us may start to promote deluxe packages to them through company’s advertising channel.

## Recommendation System

We have built and train both the neighbour-based and model-based collaborative filtering model according to the step mentioned in the methodology section. The following figure show the result of both models when given a single row of customer’s detail to demonstrate how the models predict the target variable and recommends a product. For the generalized kNN classifier, we set k = 9 because 1) we want k to be as high as possible to avoid overfitting 2) k higher than 9 slows down the algorithm with no changes to the performance because the influence of the data point beyond k = 9 on the predicted response is low due to low similarity and KD trees become slower as the k increases (scikit-learn, 2022).

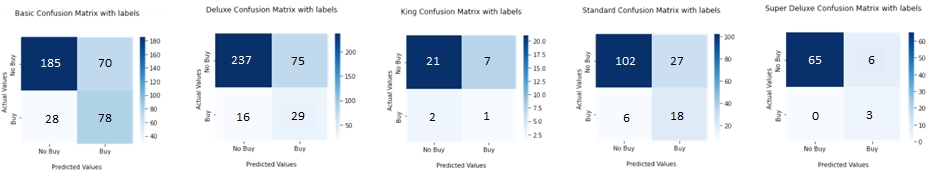
**Figure 4.5 Results of Neighbourhood and Model-based Recommender Systems**

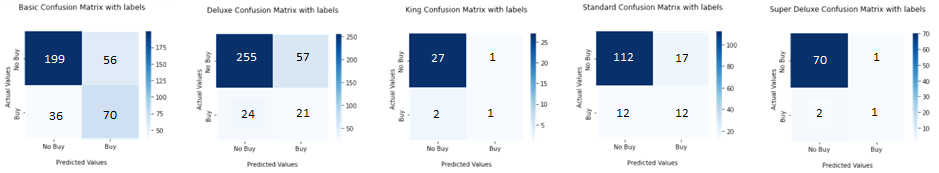
We can see that when user’s detail is inputted into the generalized kNN classifier, it output the weighted vote from the top k similar user for each product. For example, 0.58 in basic\_response means after weighted by similarity, 58% of the vote are for buying the *basic* package target variable. Then, based on these values, it recommends the deluxe package because it has the highest vote. The predicted class for each target variable is the class with most vote. For example, because 1 is the majority class in the b*asic* package target variable, the predicted class is 1 which mean the model predict that the customer will buy the package.

We can see that when user’s detail is input into the model-based LR recommender, it output the probability of buying each product. For example, the value 0.302 of basic\_prob is the probability that the input user will buy the basic product. Then, based on these values, it recommends the super deluxe package because it has the highest probability of being like/bought. The predicted class for the binary target variable is the class with highest probability. For example, because the probability of the *basic* target variable equals 1 is 0.302, the predicted class is 0 which mean that the model predicts that the customer will not buy the basic package.

## Recommendation System Evaluation

The confusion matrix is a graphical representation tool that is used to evaluate the output of a classifier. Columns contains predicted outcome (buy/not buy), rows contain the real outcome (buy/not buy) and cell denotes the number of samples predicted. Figure 4.6 and Figure 4.7 shows that the prediction outcome is similar between Neighbour-based and Model-based.

**Figure 4.6 Confusion Matrix of Neighbour-based Model**

**Figure 4.7 Confusion Matrix of Model-based Model**

Similarly, neighbour-based and model-based collaborative filtering also sharing similar model performance. However, by using one tail sample t-test (based on bootstrap sample), the accuracy of Model-based is significantly better than Neighbour-based.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Time Taken | Accuracy | Precision | Recall | F1-Score | Sig. Better |
| Neighbour-based | 4.331 s | 78% | 68% | 75% | 69% |  |
| Model-based | 0.045 s | 79% | 68% | 68% | 67% | \* |

**Table 4.4 Model Performance Comparison for Memory-based and Model-based Collaborative Filtering**

# DISCUSSION

We started this project with the following objective: 1. To determine significant features for package selection 2. To identify customer segment to target 3. To develop a package recommender to suggest tour packages to be pitched. 4. To assess the package recommender’s performance.

For feature for package selection, we found that Age, Occupation, MaritalStatus, MonthlyIncome, NumberOfPersonVisiting, PreferredPropertyStar, NumberOfTrips, Passport, Designation, CityTier, TypeofContact, NumberOfFollowups, DurationOfPitch and PitchSatisfactionScore are important.

For customer segment to target, demographically, unmarried individuals who are also business owners should be a high priority cluster to promote bigger ticket size tour packages for higher profitability. Behaviorally, Visit with Us must focus on targeting advertising material to bargain hunters as well as corporate clients to increase its sales volume. In terms of psychography, the company can expand its target market to lower budget travelers to increase its brand awareness and promote customer acquisition.

We developed two models that make recommendations based on user details. We found that the model-based recommender system performed statistically better than the memory-based. The model-based recommender also takes only 0.045 seconds to make a prediction which is 90 times faster than neighbour-based model. Therefore, the model-based recommender should be used for the tour operator, Visit with Us.

For future work, techniques apart from kNN and LR can be explored to assess the performance and accuracy for building a recommender system. Additionally, information on recurring customers can be analyzed to identify key features and understand customer behavior for repeating purchases or extending memberships. Diverse and abundant data are key to building more precise recommender systems and performing market targeting.

# REFERENCES

Aditya, P. H., Budi, I., & Munajat, Q. (2016). A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for E-commerce in Indonesia: A case study PT X. *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 303–308. <https://doi.org/10.1109/ICACSIS.2016.7872755>

Aggarwal, C. C. (2016). Recommender Systems The Textbook. In *Recommender Systems* (Vol. 39, Issue 4).

Al-bashiri, H., Abdulgabber, M. A., Romli, A., & Kahtan, H. (2018). An improved memory-based collaborative filtering method based on the TOPSIS technique. *PLOS ONE*, *13*(10), e0204434. <https://doi.org/10.1371/journal.pone.0204434>

APELO Consulting. (2022). *Market segmentation: Benefits, Types & Methods*. <https://apeloconsulting.com/market-segmentation-benefits-types-methods/>

Carraro, D., & Bridge, D. (2022). A sampling approach to Debiasing the offline evaluation of recommender systems. *Journal of Intelligent Information Systems*, *58*(2), 311–336. <https://doi.org/10.1007/s10844-021-00651-y>

Corporate Finance Institute. (2022). *Market Segmentation and Targeting*. <https://corporatefinanceinstitute.com/resources/knowledge/strategy/market-segmentation-and-targeting/>

D’Urso, P., de Giovanni, L., Disegna, M., Massari, R., & Vitale, V. (2021). A Tourist Segmentation Based on Motivation, Satisfaction and Prior Knowledge with a Socio-Economic Profiling: A Clustering Approach with Mixed Information. *Social Indicators Research*, *154*(1), 335–360. <https://doi.org/10.1007/s11205-020-02537-y>

Funk, S. (2006). *Netflix Update: Try This at Home*. <https://sifter.org/~simon/journal/20061211.html>

Gebremeskel, G. G., & de Vries, A. P. (2016). *Recommender systems evaluations: Offline, online, time and A/A test*.

Houshmand‐Nanehkaran, F., Lajevardi, S. M., & Mahlouji‐Bidgholi, M. (2022). Optimization of fuzzy similarity by genetic algorithm in user‐based collaborative filtering recommender systems. *Expert Systems*, *39*(4). <https://doi.org/10.1111/exsy.12893>

IBM. (2020). *Machine Learning*. <https://www.ibm.com/cloud/learn/machine-learning>

Kim, K. S., Chang, D. S., & Choi, Y. S. (2019). Boosting Memory-Based Collaborative Filtering Using Content-Metadata. *Symmetry*, *11*(4), 561. <https://doi.org/10.3390/sym11040561>

Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics*, *11*(1), 141. <https://doi.org/10.3390/electronics11010141>

Koren, Y. (2008). Factorization meets the neighborhood. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, 426. <https://doi.org/10.1145/1401890.1401944>

Lima, G. R., Mello, C. E., Lyra, A., & Zimbrao, G. (2020). Applying landmarks to enhance memory-based collaborative filtering. *Information Sciences*, *513*, 412–428. <https://doi.org/10.1016/j.ins.2019.10.041>

Ortega, F., Lara-Cabrera, R., González-Prieto, Á., & Bobadilla, J. (2021). Providing reliability in recommender systems through Bernoulli Matrix Factorization. *Information Sciences*, *553*, 110–128. <https://doi.org/10.1016/j.ins.2020.12.001>

Kaggle. (2022). TourPackagePrediction. <https://www.kaggle.com/datasets/sanamps/tourpackageprediction>

scikit-learn. (2022). *1.6. Nearest Neighbors*. <https://scikit-learn.org/stable/modules/neighbors.html#unsupervised-neighbors>

Sekhar, P. R., & Sujatha, B. (2020). A Literature Review on Feature Selection using Evolutionary Algorithms. *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, 1–8. <https://doi.org/10.1109/ICSSS49621.2020.9202257>

World Tourism Organization. (2022). *INTERNATIONAL TOURISM AND COVID-19*. <https://www.unwto.org/tourism-data/international-tourism-and-covid-19>