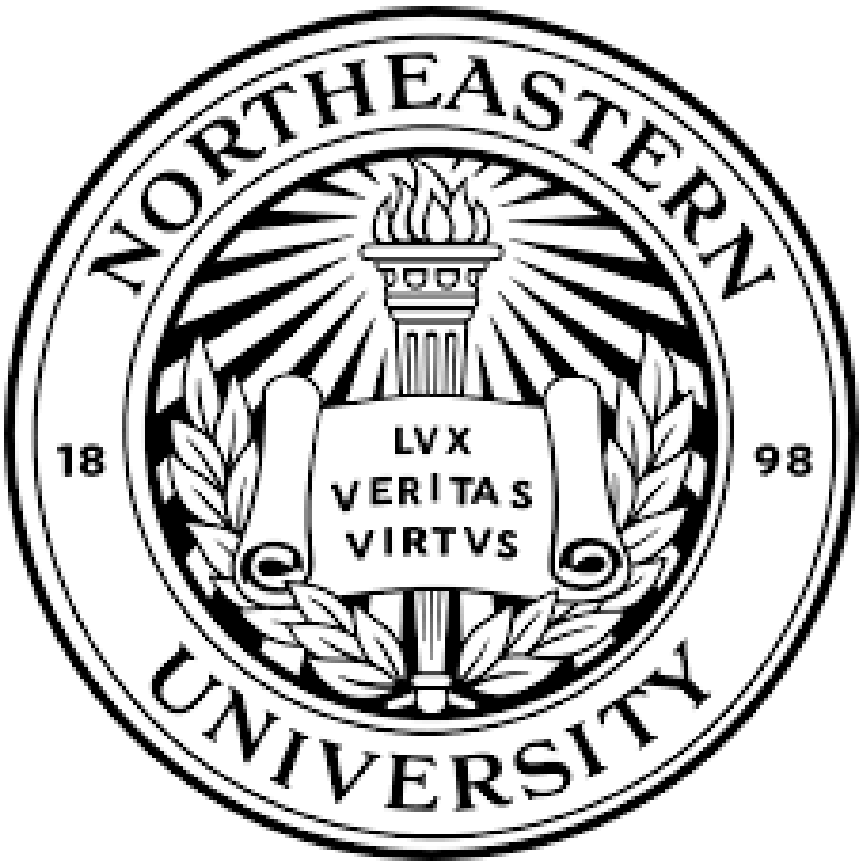
Data Science Engineering Methods and Tools

Tourism Growth Prediction



Submitted by –
Mandar Deshmukh (002194579)
Rituja Lolam (002146428)

Table of Contents

1. Introduction

- **Tourism** is an act and process of staying away from our own house in pursuit of relaxation, business, leisure, fun, sightseeing etc.
- **Tourism** boosts the revenue of the economy, creates thousands of jobs, develops the infrastructures of a country, and plants a sense of cultural exchange between foreigners and citizens.

A reputed big tourism company is facing a problem and unable to solve it. They don't have any idea about what are the important key points to track and resolve it. So, we will solve the business problem using various Machine Learning techniques

Dataset Link: https://figshare.com/articles/dataset/Tourism_csv/19350668

1.1 Brief introduction about the problem statement

Problem Statement:

A reputed tourism company is planning to launch a long-term travel package. The Product Manager has access to the existing customers' data and information. He wishes to analyse the trend of existing customers to figure out which customer is going to purchase the long-term travel package.

Issues faced:

- They don't know who are the primary customers helping to run their business.
- They don't have the right plan to increase the business revenue.
- New plan which they are going to introduce will attract customers or not.
- They are lagging business key ideas.
- They wanted to predict whether any customer will opt for long-term package providing their details.
- Here the predictor variable is **ProdTaken**. It says whether that customer has taken that long-term package or not.

1.2 Objective

- **Objective** of this project is to integrate all the learning skills and to find out the solution for real world business problem.
- This project will cover Statistics, EDA, Model Building, Testing, Model Evaluation etc.
- Solving this problem and providing recommendations will help the company to grow.

2. EDA and Business Implication

a) Understanding the data

- Customers tour related information, working details, their enquiry, follow ups, etc are collected.
- Time frequency is not available with us because there is no date/time field information in our data set.
- Features given are 'CustomerID', 'ProdTaken', 'Age', 'CityTier', 'DurationOfPitch', 'NumberOfPersonVisited', 'NumberOfFollowups', 'PreferredPropertyStar','NumberOfTrips', 'Passport', 'PitchSatisfactionScore', 'OwnCar', 'NumberOfChildrenVisited', 'MonthlyIncome', 'PreferredLoginDevice', 'Occupation', 'Gender', 'ProductPitched', 'MaritalStatus', 'Designation'
- It contains dependent variable **(ProdTaken)** so it is a supervised learning method.

b) Visual inspection of data (rows, columns, descriptive details)

- The given data has been imported into out working frame using pandas library. The data set contains 20 different features of customer and their travel information.
- Number of columns (i.e., Features) present in our data set is **20**.
- Descriptive statistics for the given data

    - **Customer ID** column contains customer id’s of 4888 members.

- **ProdTaken** field contains binary values 0 and 1. Most number of occurrences are 0 with count 3968.
- **Age** is a continuous variable having age of customers from range 18 to 61. Mean value is 37.622 and median is 36.
- **CityTier** has the city information of customer. City Tier 1 indicates it is a metropolitan city. Most of the customers about 3190 customers live in Tier 1 city.
- **Duration of Pitch** is a continuous nature having range from 5 to 127. Mean value is 15.49 and median is 13.
- Number of Persons Visited(**NumberOfPersonVisited**) with customers is a discrete variable having values from 1 to 5
- Number of followups(**NumberOfFollowups**) done by sales person after sales pitch is from 1 to 6.
- **NumberOfTrips** is a no of trips went by each customer. Least trip count went by a customer is 1 and max count is 22.
- **PreferredPropertyStar** is accommodation liked by customer to stay. Customer likely to book 3 star to 5 star hotel.
- **NumberOfChildrenVisited** with customers is a discrete variable having values from 1 to 3. Nearly half of the customers come with 2 childrens.
- **PitchSatisfactionScore** is score rating given by customer for Product pitching done by sales person.
- **OwnCar**, **Passport** are binary variables having values as 0 and 1. They just tell whether customer has own car and passport.
- **MonthlyIncome** is the income of each customer. A customer has least salary of 1000Rs and max salary up to 98678Rs
- **PreferredLoginDevice** is customer's preferred login device for enquiry for tour booking. 3444 customers had done self-Inquiry.
- **Occupation** is a type of work doing by customer. It has Salaried, Free Lancer, Small Business, Large Business occupation types.
- Database have equal customers in **gender** wise.
- **ProductPitched** is the product pitched by customers. It has 5 values Multi, Super Deluxe, Standard, King, Deluxe.
- Database has more **married** customers data. 2340 are married

- More executives are present in our data set. 1842 customers are executives. 5 different **designation** values are present in our data Executive, Manager, Senior Manager, AVP, VP.

3. Model building

3.1) Clear on why was a particular model(s) chosen

- Our problem is a binary classification problem so we built classification models
- Model(s) for machine learning algorithm is chosen based on same factors:
  - The size, quality, and nature of data.
  - The available computational time.
  - The urgency of the task.
- If Explanation is needed for the model is needed then we can use Regression and Decision Tree models.
- Feature importance is easily identified using the Decision Tree and Regression model
- If accuracy and speed is only concern then we can go for SVM, Random Forest Models.
- If Data is too large then we can go for Naïve-Bayes model. Naïve bayes also works well incase of more features are categorical
- Boosting models are slower because they learn from the previous model built.
- Our concern is to have accuracy, recall, precision, F1 score, AUC score values for the model. So, we built all the models and compared which is better.
- Various models built using Logistic Regression, Tree based models, Ensemble methods, Linear Discriminant analysis, K Nearest Neighbour, Bayes Theorem.
- As we don't have separate test data to validate our model. So, we will split the whole data set into train and test.
- We will split at a ratio of 70:30 train and test.
- This can be done using sklearn.preprocessing library and train_test_split function. We pass variables X, y, test_size as arguments to this function.
- Test size is given as 0.3 to take 30% of data from the whole data

3.2 Effort to improve model performance.

- Model tuning, hyper parameter optimization was done to improve model performance.
- All the models were tuned to provide better results.
- GridSearchCV function from sklearn library helps to find out the better value for parameters present in the model.
- Cross validation also taken into account while doing GridSearchCV to prevent model over fitting.

Models Performance

Decision Tree Model:

- **Decision tree model is pruned to achieve better results and better prediction.**
- **After pruning best parameters identified are** {'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 5, 'min_samples_split': 15}.
- This model is now made fit to predict from overfitting.
- After accuracy score of train and test data are 83 and 82%.
- Average F1 score value seen in this model.
- AUC ROC score for train and test data are 84 and 78%

ANN :

- Default values of hyper parameters were used to built the model.
- This model shows an accuracy of 81 % in both train and test data.
- Poor Recall, Precision scores were observed with this model.
- No 1's prediction done by this model.
- Hyper parameter optimization needs to be done to achieve best results.
- ANN doesn't seems to perform well after optimization.

Random Forest

- This is also a tree-based model helps in predicting the target variable by building large number of decision trees.

- This model also shows an accuracy of 100% results with train data and 90% with test. This model is an overfitting model.
- As Random Forest model was not pruned properly so it leaded to model overfitting.
- We can see a large difference between Random Forest train and Test models. 100% accuracy in training data and 90% accuracy in testing data.
- This model is pruned to get better parameters.
- GridSearchCV function is used to find out the best parameter after many ite rations.
- Best parameters identified are {'max_depth': 10,'max_features': 13, 'min_sa mples_leaf': 5, 'min_samples_split': 15, 'n_estimators': 701}
- After pruning this model gives the better results.
- Accuracy score of 92% and 86% with train and test data.
- Better precision and recall values seen in this model over other models.
- Moderate level recall and precision scores are seen in this model for test da ta.
- Best AUC score is seen in model at 98% and 91% train and test data.

Bagging

- Bagging model is tuned to get the best hyper parameters.
- Best parameters identified are {'max_features': 13, 'max_samples': 100, 'n_estimators': 100}.
- Bagging model gives the accuracy of 83% train and 82% test.
- It gives very poor performance metrics recall value. But precision value is seen better in this model.
- AUC Score is 86 and 80% with train and test data.

GradientBoosting

- **Gradient Boosting** model is tuned to get the best hyper parameters.
- Best parameters identified are {'learning_rate': 0.01, 'max_depth': 9, 'n_esti mators': 1000, 'subsample': 1.0}
- Gradient model gives the accuracy of 100% train and 93% test.

- It seems as an unfit model
- 100% results seen in all performance metrics. This makes model overfitting

AdaBoosting

- **Ada Boosting** model is tuned to get the best hyper parameters.
- **B**est parameters identified are {'learning_rate': 0.001, 'n_estimators': 10}
- Ada Boosting model gives the accuracy of 100% train and 89% test.
- This model is overfitting model
- 100% results seen in all performance metrics. This makes model overfitting

KNN:

- Next we tried to find out the optimal k value for our data set for developing KNN model. We choose multiple K values and predict the results and error value.
- We find out the best k value using Elbow Method.
- Plotted K vs WSS(within sum of square) values against each other.
- Choosen the best K – Value which has minimum error value. **K=11**
- Overall accuracy seen at 81% with test and train data.
- But poor recall and precision, f1score observed with this model.
- Less AUC ROC score also seen for this model at 62% for test data.

SVC:

- After model tuning accuracy and other performance metrices improved mu ch better than before tuning.
- Best parameters found for SVM algorithm after tuning are {'C': 10, 'gamma': 'auto', 'kernel': 'rbf'}
- Model accuracy seen at 100% for the train data and 89% for test data.
- 100% train data results seen in all performance metrics. This makes model overfitting
- AUC ROC score also seems overfitting.

Naïve Bayes

- When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.
- Naïve Bayes model before and after tuning produces the same accuracy results with 82%.
- Better recall, precision, f1 scores are seen than SVM model from Naïve Bayes model.
- Not much hyper parameters are there for this model.
- So, tuning won't improve this model accuracy a lot.
- This model seems to perform at moderate level.

Logit Model

- As we seen in previous model many insignificant values are seen in model b uilding. Also, we checked whether any multicollinearity problem in model b uilding.-

VIF:

A variance inflation factor(VIF) detects multicollinearity in regression analysis. Mul ticollinearity is when there's correlation between predictors (i.e. independent vari ables) in a model; it's presence can adversely affect your regression results. The VI F estimates how much the variance of a regression coefficient is inflated due to m ulticollinearity in the model.

- Normally VIF value higher than 10 have more multicollinearity between the m.
- Those can be removed while model building.

- So MonthlyIncome, Salaried, Age, Small_Business, NumberOfPersonVisited, PreferredPropertyStar, Designation, NumberOfFollowups variables can be r emoved the data.
- Log2 model was built with all significant variables.
- Next is to find out the best cut off point to separate the two classes.
- ROC_CURVE function best cut off point identified is **0.1823**

- Train and test data are validated with this optimized model. Accuracy score i s at 70%
- Higher recall value is seen in this model other than all models.
- F1 score is at average of 50%
- Passport, Single, married, Large Business, City Tier are the positive variables help in predicting the target class.
- ProductPitched, LoginType are the variables have negative impact on target variable.

	AUC	Accuracy Score	Recall Score	Precision Score	F1 Score
Logistic_Train	73.47	82.78	14.13	71.65	23.61
LDA1_Train	81.06	84.39	31.52	68.58	43.19
Bagging_Train	85.99	83.60	13.98	92.78	24.29
Decision Tree Tuned_Train	84.04	85.71	38.20	73.00	50.15
ANN Tuned_Train	64.11	81.18	0.00	0.00	0.00
RandomForestTuned_Train	98.14	92.17	62.27	94.13	74.95
KNN_Tuned_Train	77.48	81.12	9.01	49.15	15.22
SVC_Tuned_Train	100.00	100.00	100.00	100.00	100.00
Naive2_Tuned_Train	77.09	83.81	30.43	64.90	41.44
LogitTunedTrain	77.34	70.44	74.00	36.00	48.00

Fig 4.2 i) Performance table of model with train data.

	AUC	Accuracy Score	Recall Score	Precision Score	F1 Score
Logistic_Test	68.22	82.21	10.51	67.44	18.18
LDA1_Test	78.62	82.75	21.74	61.86	32.17
Bagging_Test	79.90	82.82	10.14	87.50	18.18
Decision Tree Tuned_Test	78.66	84.19	32.61	66.18	43.69
ANN Tuned_Test	61.00	81.19	0.00	0.00	0.00
RandomForestTuned_Test	91.05	86.43	38.77	78.10	51.82
KNN_Tuned_Test	62.97	80.50	5.80	38.10	10.06
SVC_Tuned_Test	88.97	89.84	49.64	93.20	64.78
Naive2_Tuned_Test	73.24	83.37	26.81	63.79	37.76
LogitTunedTest	75.87	70.06	70.00	35.00	47.00

Fig 4.2 ii) Performance table of model with test data.

4. Model validation

4.1 How was the model validated? Just accuracy, or anything else too?

- All models are tested with test data.
- Test data is formed by random 30% of the input sample data.
- Performance metrics like accuracy, precision, f1-score, recall value, AUC score all are taken into account while building all the models.
- Classification report, Confusion matrix gives the result of model's performance.
- Only accuracy was not taken into for validation.

Classification Report

The classification report visualizer displays the precision, recall, F1, and support scores for the model.

Precision: It is defined as the ratio of true positives to the sum of a true positive and false positive. It is called accuracy of positive predictions.

Recall: It is defined as the ratio of true positives to the sum of true positives and false negatives. True positives that were correctly identified.

F1 Score: The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

Support**:** Support is the number of actual occurrences of the class in the specified dataset.

Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

• true positives (TP): These are cases in which we predicted yes.

• true negatives (TN): We predicted no, and they will not opt for package.

• false positives (FP): We predicted yes, but they don't opt for package. (Also known as a "Type I error.")

• false negatives (FN): We predicted no, but they do subscribe for package. (Also known as a "Type II error.")

AUC-ROC:

AUC - ROC curve is a performance measurement of prediction for classification problem.

ROC is a probability curve and AUC represent degree or measure of separability.

It tells how much model is capable of distinguishing between classes.

Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

Accuracy Score

It is the ratio of number of correct predictions to the total number of input samples.

- Many models are built with the train data and tested with the test data.
- Their performance metrics are compared to check which model is doing good.
- Fit function is used to fit the train data with the model.
- Predict function is used to predict the output of the test data.
- PrintModelReport function is made to print the performance of the individual model.
- Performance Table is made to check the performance metric of the models.
- Performance Table is made with AUC ROC Score, Accuracy Score, Recall score, Precision Score, F1 Score.

5) Final interpretation / recommendation - Very clear and crisp on what recommendations do you want to give to the management / client

5.1) Final Model Interpretation:

- Among all the models built RandomForest classifier and Logit model seems to be better.
- As seen from accuracy, AUC score, precision, recall value, f1 score metrics Random Forest model seems to be better than Logit model.
- Logit model is better in Recall metric than Random Forest model.
- Recall value metric is most important metric where it says how many true positives are identified correctly.
- When taking recall metric into account Logit model seems too overall good.
- Recall metric is where customers opting for package and we need to predict them correctly.
- If we predict them correctly then we can go with the tour package and make the business grow better.
- True positives identified by the model in test data is 835
- True negatives identified by the model in test data is 192
- Around 75% of the data fits the ROC AUC Curve

- Passport, Single, married, Large Business, City Tier are the positive variables help in predicting the target class.
- ProductPitched, LoginType are the variables have negative impact on target variable.
- 70% of the Customer who is having passport, single/unmarried doing Large Business and from less city tier will take the tour package.
- Overall 70% prediction is a good model in Holiday/Tour industry. But we need to improve it better.

5.2) Conclusion-

Amongst all the models trained, we observed that Randomforest Classifier and Logit model performed better.

For Randomforest Classifier-
Accuracy, AUC score, precision, recall value, f1 score were better.

For Logit Model-
Recall metric was better.

Around 75% of the data fits the ROC AUC curve. Also, True Positives and Negatives identified by the model are 835 and 192 respectively.

Passport, Single, Married, Large Business, City Tier are the positive variables and help in predicting target variable.

Product Pitched and LoginType are negative variables and affect the target variable negatively.

5.3) Recommendations to the management

- ✓ **Pitch satisfaction needs to be improved a lot**
- ✓ **More follow ups and offers can be given to 2nd, 3rd tier customers**
- ✓ **Family, couple packages can be introduced to cover married customers**
- ✓ **Prices for Product Pitched can be reduced a little.**
- ✓ **Recommending to take passport for all customers.**
- ✓ **Outliers and missing values to be taken care.**
- ✓ **Analyze the reviews of all customers and notify the improvement made on negative part.**