



VISITWITHUS LLC

data analytics – travel package purchase prediction

by Marcelo Moraes

COMPANY'S FUTURE VISION



ENABLE AND ESTABLISH A VIABLE
BUSINESS MODEL TO EXPAND THE
CUSTOMER BASE



UNDERSTAND THE NECESSARY
CHANGES IN THE BUSINESS TO
BENEFIT TOURISM SECTOR



MAKE THE MARKETING
EXPENDITURE MORE EFFICIENT



DATA ANALYTICS FOCUS

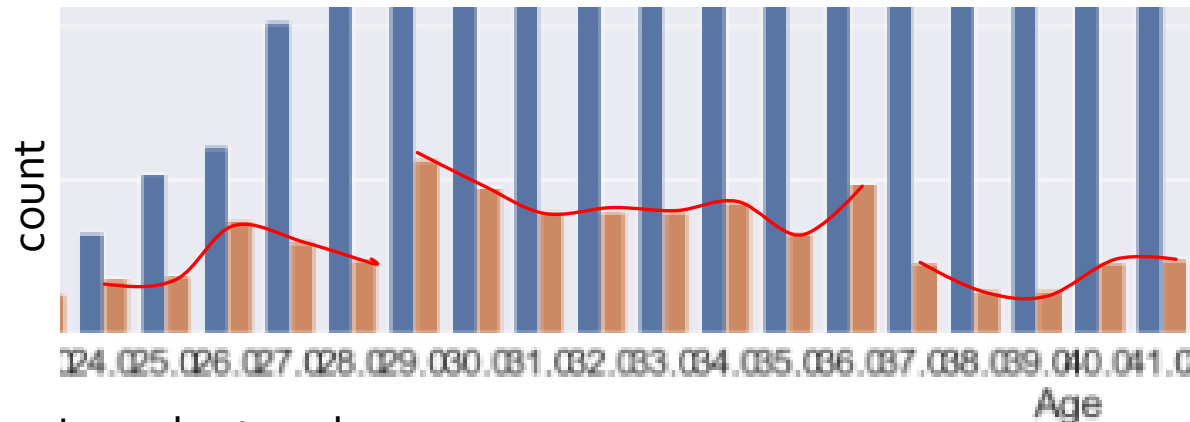
- PROVIDE INSIGHTS TO SUPPORT INTRODUCTION OF NEW OFFERING OF PACKAGES
- PROVIDE A BETTER UNDERSTANDING ABOUT THE CUSTOMER PROFILE AND TARGET GROUP.
- BUILD A MODEL TO PREDICT BUYING INTENT

| VARIABLES | DESCRIPTION |
|-------------------------|---|
| CustomerID | Unique customer ID |
| ProdTaken | Product taken flag |
| Age | Age of customer |
| TypeofContact | How customer was contacted (Company Invited or Self Inquiry) |
| CityTier | City tier |
| DurationOfPitch | Duration of pitch by a sales man to customer |
| Occupation | Occupation of customer |
| Gender | Gender of customer |
| NumberOfPersonVisited | Total number of person came with customer |
| NumberOfFollowups | Total number of follow up has been done by sales person after sales pitch |
| ProductPitched | Product pitched by sales person |
| PreferredPropertyStar | Preferred hotel property rating by customer |
| MaritalStatus | Marital status of customer |
| NumberOfTrips | Average number of trip in a year by customer |
| Passport | Customer passport flag |
| PitchSatisfactionScore | Sales pitch satisfactory score |
| OwnCar | Customers owns a car flag |
| NumberOfChildrenVisited | Total number of children visit with customer |
| Designation | Designation of customer in current organization |
| MonthlyIncome | Gross monthly income of customer |



**DATA
INFORMATION**
(88559 data points)
(1.1% missing values)

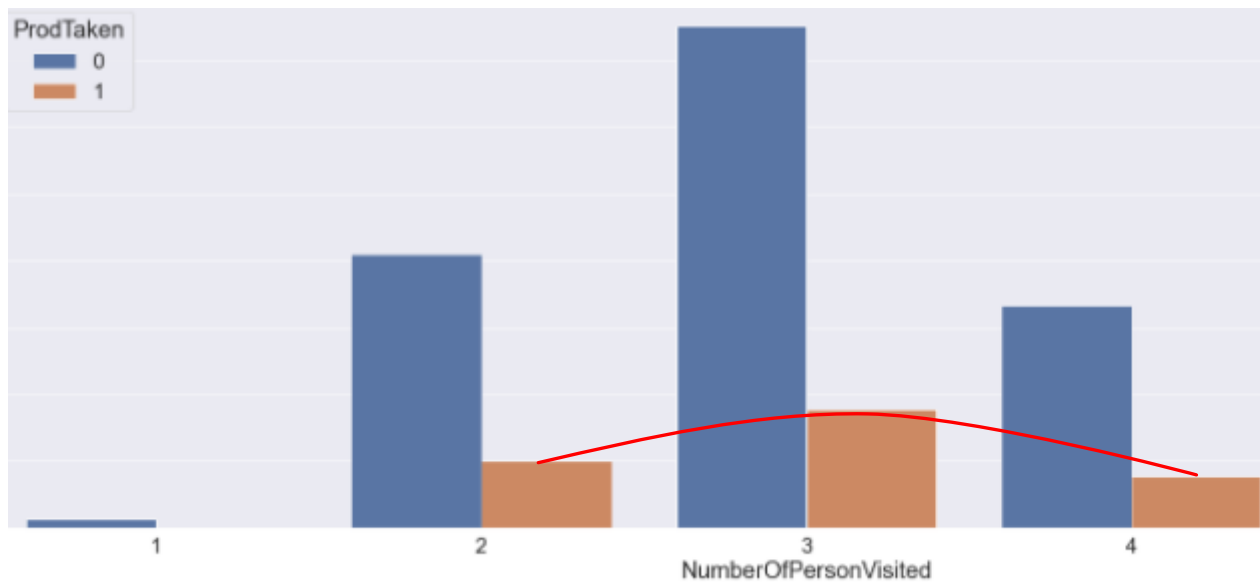
- Salaried and small business owners represent 91% of customer base
- 47.9% of customers are married
- 71.7% of the customer base are Executives or Managers
- Preferred destinations are cities Tier 1
- 3 stars properties represented 62% of total choices regardless of the customers designation
- Ages between 29 and 36 outstand in number of packages purchased



* Red line showing sales trend

CUSTOMER
PROFILE
IDENTIFICATION

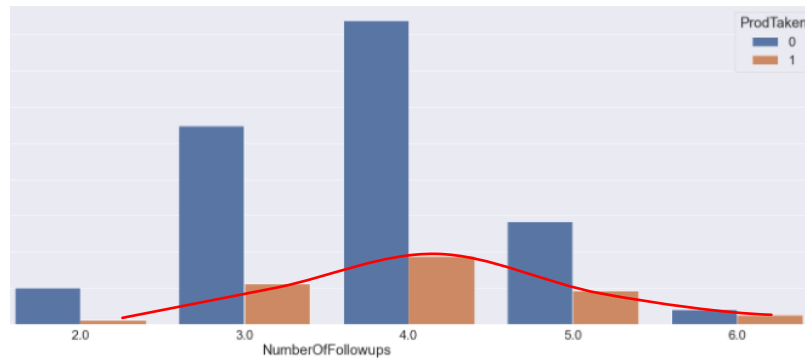
- 71% of the type of contact was self enquiry
- Numbers of people visiting sales point also shows a hidden pattern, when number of people visiting sales point is between 2 and 3, higher are the chances to turn into sales.



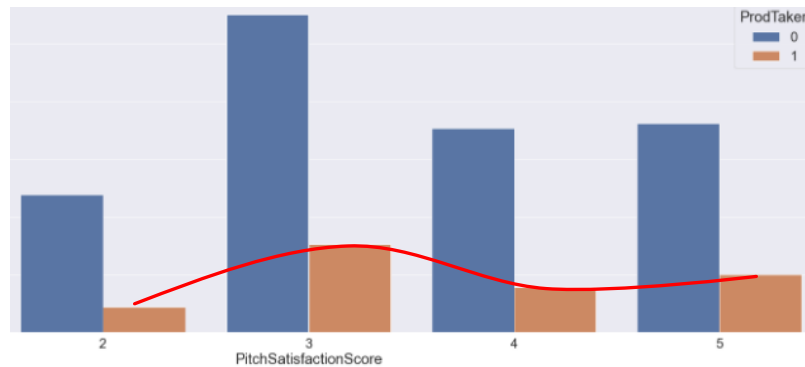
BUSINESS
INSIGHTS

* Red line showing sales trend

- Numbers of follow-ups , minimum 3 and maximum 5 shows better chances to turn into sales.



- Pitch Satisfaction score also shows scores between 3 and 5 boost sales



* Red line showing sales trend

BUSINESS
INSIGHTS

- 14 machine learning models were used in attempt to find the best predictor of a buyer intent.

| | Model | Train_Precision | Test_Precision | Delta |
|----|--|-----------------|----------------|----------|
| 0 | Decision Tree | 1.000000 | 0.656140 | 0.343860 |
| 1 | Tuned Decision Tree | 0.702247 | 0.610294 | 0.091953 |
| 2 | Bagging Classifier | 1.000000 | 0.834225 | 0.165775 |
| 3 | Tuned Bagging Classifier | 1.000000 | 0.859459 | 0.140541 |
| 4 | Random Forest | 1.000000 | 0.877698 | 0.122302 |
| 5 | Tuned Random Forest | 1.000000 | 0.868966 | 0.131034 |
| 6 | Adaboost Classifier | 0.721362 | 0.601626 | 0.119736 |
| 7 | Gradient Boosting Classifier | 0.880208 | 0.712000 | 0.168208 |
| 8 | XGBoost classifier | 1.000000 | 0.827586 | 0.172414 |
| 9 | Tuned Adaboost Classifier | 0.978648 | 0.743455 | 0.235192 |
| 10 | Gradient Boosting Classifier with Adaboost | 0.888000 | 0.706897 | 0.181103 |
| 11 | Tuned Gradient Boosting Classifier | 0.949002 | 0.753247 | 0.195755 |
| 12 | Tuned XGBoost Classifier | 0.994231 | 0.809211 | 0.185020 |
| 13 | Stacking Regressor | 0.987200 | 0.751020 | 0.236180 |

MODEL OVERVIEW

- Precision was the metric adopted to evaluate performance of the model:
 - Predict a customer will purchase a tourism package but customer does NOT purchase it also called False Positive was targeted to minimize such error in order to lower the marketing expenditure in customers that would not purchase the offered packaged at all.
- The main focus of the model was to maximize Precision, consequently lowering False Positive rates.



MODEL
METRICS

- The 3 models below were the ones which performed best among all others:

| | Model | Train_Precision | Test_Precision | Delta |
|----|------------------------------------|-----------------|----------------|----------|
| 9 | Tuned Adaboost Classifier | 0.978648 | 0.743455 | 0.235192 |
| 11 | Tuned Gradient Boosting Classifier | 0.949002 | 0.753247 | 0.195755 |
| 13 | Stacking Regressor | 0.987200 | 0.751020 | 0.236180 |

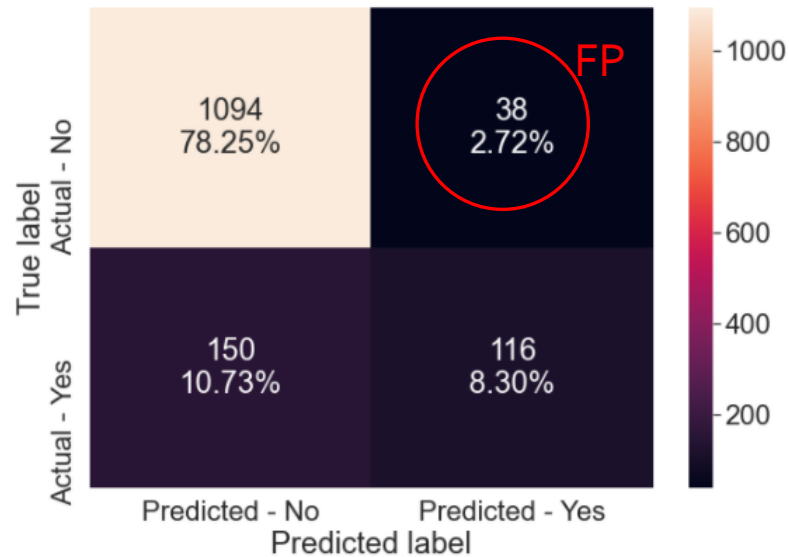
- Tuned Gradient Boosting Classifier was selected as the best predictor with the highest precision value on test data.

| | Model | Train_Precision | Test_Precision |
|----|------------------------------------|-----------------|----------------|
| 11 | Tuned Gradient Boosting Classifier | 0.949002 | 0.753247 |

- Main reasons for the model selection:
 - First, GBC tuned had the highest precision value
 - Second, GBC tuned presented lower chances to overfitting since the precision in the train set was again the lowest value among all 3 models.

MODEL
SELECTION

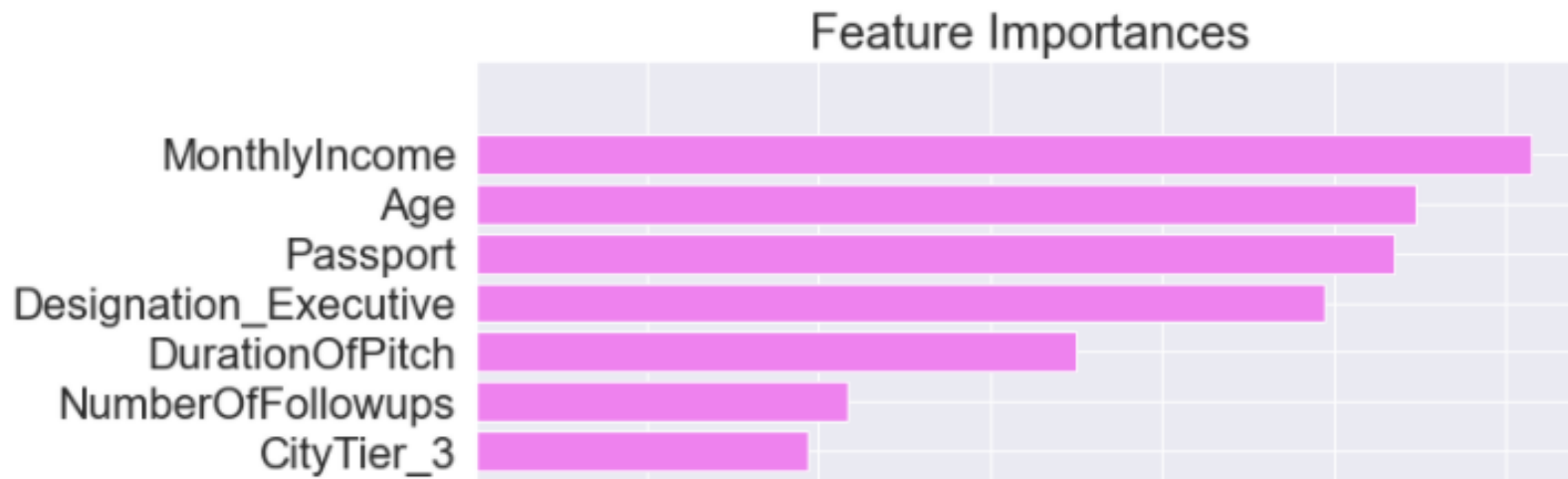
- Confusion Matrix, main points overview



MODEL OVERVIEW

- The winner model (GBC Tuned) conclusion:
 - Model was capable to minimize false positive error by 2.72%. In other words, the model was capable to correctly predict 86.55% of customers would purchase or not the offered package and it would wrongly predict (FP) 2.72% of the cases.
 - The remaining 10.73% (FN) model predicted a customer would not purchase, but in fact they did.

- Feature importance shows what were the most important features in ascending order the model used and get the obtained results.



MODEL
OVERVIEW

- An interesting factor is that the first 4 features in this case are customer related features and the following 3 business related.

- Define as a target group Salaried and small business owners
- 47.9% of customers being married, may trigger additional services like childcare, summer camp or recreation per age.
- 71.7% of the customer base being Executives or Managers. Research should be done to identify main keys of communicate with such public
- Preferred destinations being cities Tier 1 give a direction to formulate business strategy more focused on increase customer experience rather than spend resources to grow the selection portfolio.
- 3 stars properties should be the target for improvements in cost, communication, technology, availability schedule, percentage of availability for the business and increased services partners.
- Exclusive marketing camping may target customer within age 29 and 36 and their respective preferences.
- 71% of the type being self enquiry shows that resources should be prioritized to focus on advertisement and targeted publicity and less company invitations.
- Different strategies and different approaches may be created based on the numbers of people visiting sales point.
- Define a threshold min/max for follow-ups should be done in order minimize losses, suggestion minimum 3 and maximum 5.
- Focus on increase Pitch Satisfaction Score by training field personal might significantly boost sales. Rates from 3 to 5 should be the target.



RECOMMENDATIONS FOR THE BUSINESS