IIT ROORKEE

Cognizance 2021

>> THE TECHNICAL EXTRAVAGANZA

BUSINESS & FINANCE

ALANKAN

Let's analyze

Team Acquisitions Inc.

Pune Institute of Computer Technology

- Maurvin Shah | cogni2005707 | maurvinshah4@gmail.com | +91 8208382431
- Aditya Khaire | cogni2005737 | adityakhaire11@gmail.com | +91 7387123852
- Sharlene Wadhwa | cogni2005721 | sharlwadhwa@gmail.com | +91 9766093436

Click on the Icons(github or drive) to access the .ipynb file (python Code)

GitHub

OR

Index

Introduction

This case study is about **Invistico Airlines** which has a growing customer base over the past few years. The strategy is helping the airline in keeping the cost of operation low and passing on the benefits to end customers. As it gives a very cost-effective price to customers, the airline wants to know whether their customers are satisfied with the services.

In this case study, we have used the dataset and:

1. Found Top factors affecting customer satisfaction.
2. Built a model to predict the customer satisfaction.
3. Calculated Overall Satisfaction Level for each passenger.
4. Found Key Inferences and Insights to increase customer contentment.

Data Description

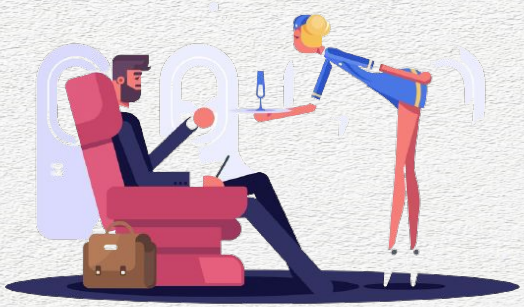
- We Use the Invistico Airline Dataset with **23 attributes**, their descriptions are given Here.

- The Dataset consists of **22 feature columns**, and **1 target** column (satisfaction).

- The Target column is a binary attribute, that means it has only two unique values i.e. ‘satisfied’ and ‘dissatisfied’

- Out of the 21 columns, **14 are survey entries** done by the passengers as a feedback.

- The passengers have rated their flight experience between **1-5**, 1 being lowest and 5 being highest, over multiple factors. (Such as Ease Of Online Booking, Seat Comfort, etc)



Data Cleaning and Preparation

(i) Removing 0 (N/A) values

- The Dataset contains exactly **129880 non-null** entries with passenger ratings over different aspects of the flight.
- Out of these, many entries are **0 (not applicable)**, we assumed them to be unfilled survey values.
- Hence, we **discard** all the rows having entries as **0** (even a single entry). This would result in a cleaner data for model building.
- After removing these values, we are left with **119611 non-null** entries.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 119611 entries, 1037 to 129879
Data columns (total 23 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   satisfaction                      119611 non-null object
 1   Gender                            119611 non-null object
 2   Customer Type                     119611 non-null object
 3   Age                               119611 non-null int64
 4   Type of Travel                    119611 non-null object
 5   Class                             119611 non-null object
 6   Flight Distance                   119611 non-null int64
 7   Seat comfort                      119611 non-null int64
 8   Departure/Arrival time convenient 119611 non-null int64
 9   Food and drink                    119611 non-null int64
 10  Gate location                     119611 non-null int64
 11  Inflight wifi service             119611 non-null int64
 12  Inflight entertainment            119611 non-null int64
 13  Online support                    119611 non-null int64
 14  Ease of Online booking            119611 non-null int64
 15  On-board service                  119611 non-null int64
 16  Leg room service                  119611 non-null int64
 17  Baggage handling                  119611 non-null int64
 18  Checkin service                   119611 non-null int64
 19  Cleanliness                       119611 non-null int64
 20  Online boarding                   119611 non-null int64
 21  Departure Delay in Minutes        119611 non-null int64
 22  Arrival Delay in Minutes          119255 non-null float64
dtypes: float64(1), int64(17), object(5)
```

(ii) Dealing with Categorical Data

- Categorical variables represent types of data which may be divided into groups.

  For Example: Gender variable can be divided into two groups, Male and Female

- We **cannot use** attributes with **object data type** for model training. Hence we **convert** the Category Variable (with 2 unique values) **into boolean Integers**.

  i.e. Gender (Male, Female) becomes Gender (1,0)

- We convert the following attributes from the given dataset to their respective boolean integer values :
  ‘satisfaction’, ‘Gender’, 'Customer Type', 'Type of Travel'

In [33]: df.head()

Out[33]:

	satisfaction	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time convenient	Food and drink	...	Online support	Ease of Online booking	On-board service	Leg room service	Baggage handling
1037	dissatisfied	Male	Loyal Customer	48	Personal Travel	Eco	4001	1	1	1	...	1	1	4	1	1
1038	dissatisfied	Male	Loyal Customer	48	Personal Travel	Eco	3980	1	1	1	...	4	4	2	3	2
1041	dissatisfied	Male	Loyal Customer	40	Personal Travel	Eco	2251	1	1	1	...	1	1	3	3	2
1043	dissatisfied	Male	Loyal Customer	46	Personal Travel	Eco	2453	1	1	1	...	5	3	1	4	1
1044	dissatisfied	Male	Loyal Customer	63	Personal Travel	Eco	2011	1	1	1	...	5	5	1	5	2

5 rows × 23 columns

In [165]: df.head()

Out[165]:

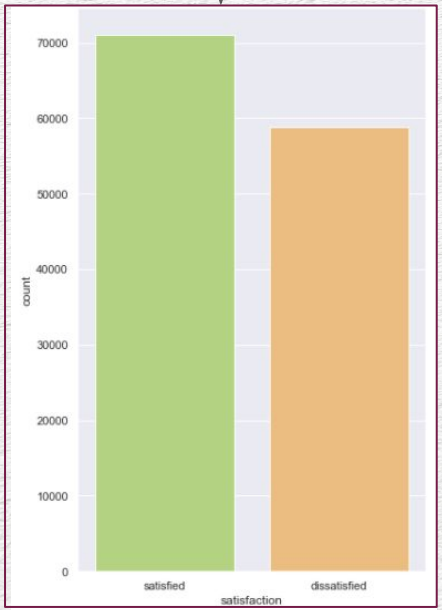
	satisfaction	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time convenient	Food and drink	...	Online support	Ease of Online booking	On-board service	Leg room service	Baggage handling
1037	0	1	1	48	1	Eco	4001	1	1	1	...	1	1	4	1	1
1038	0	1	1	48	1	Eco	3980	1	1	1	...	4	4	2	3	2
1041	0	1	1	40	1	Eco	2251	1	1	1	...	1	1	3	3	2
1043	0	1	1	46	1	Eco	2453	1	1	1	...	5	3	1	4	1
1044	0	1	1	63	1	Eco	2011	1	1	1	...	5	5	1	5	2

5 rows × 23 columns

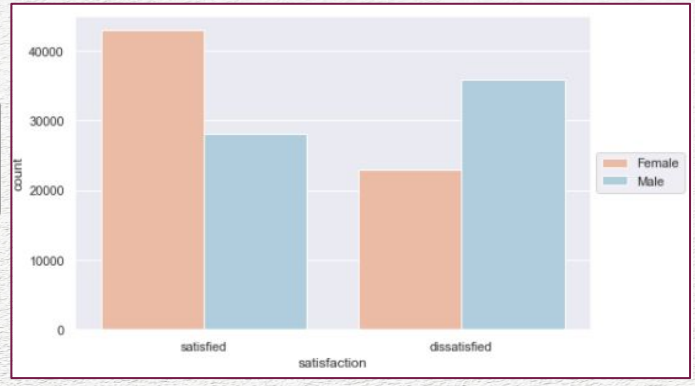
Categorical data converted to boolean integers

Exploratory Data Analysis

- When we compare the total satisfied and dissatisfied customers, we cannot infer much.

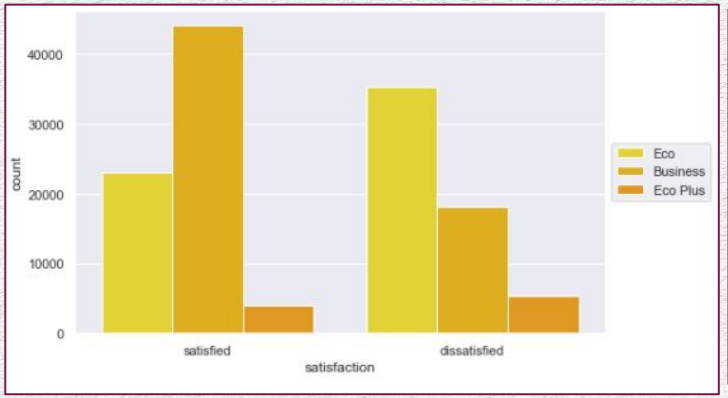


There is a higher Male dissatisfaction than Female

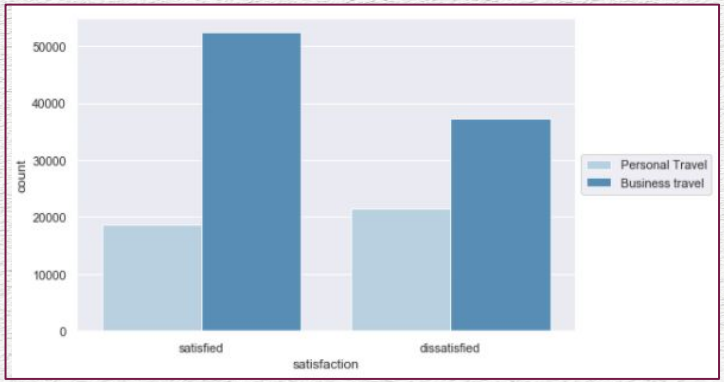


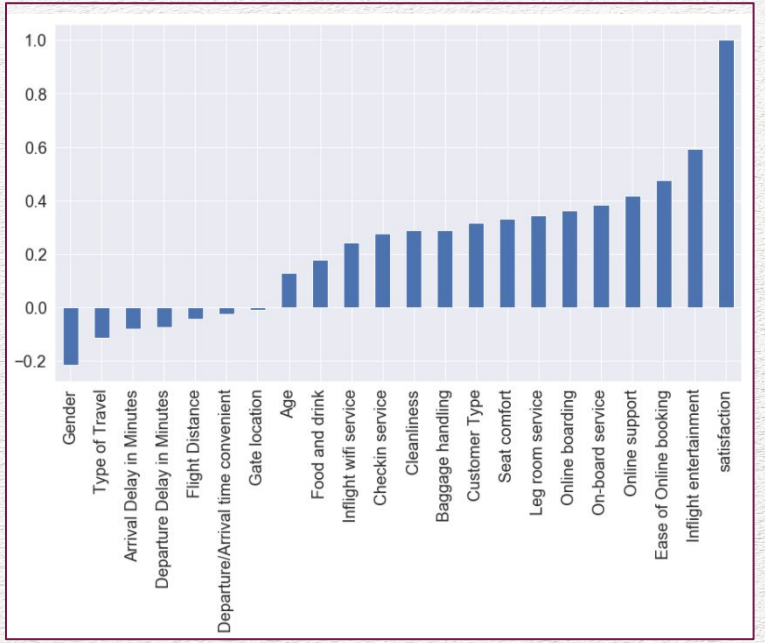
Hence we plot distribution of satisfied and dissatisfied customers over the categorical parameters. And Infer the following.

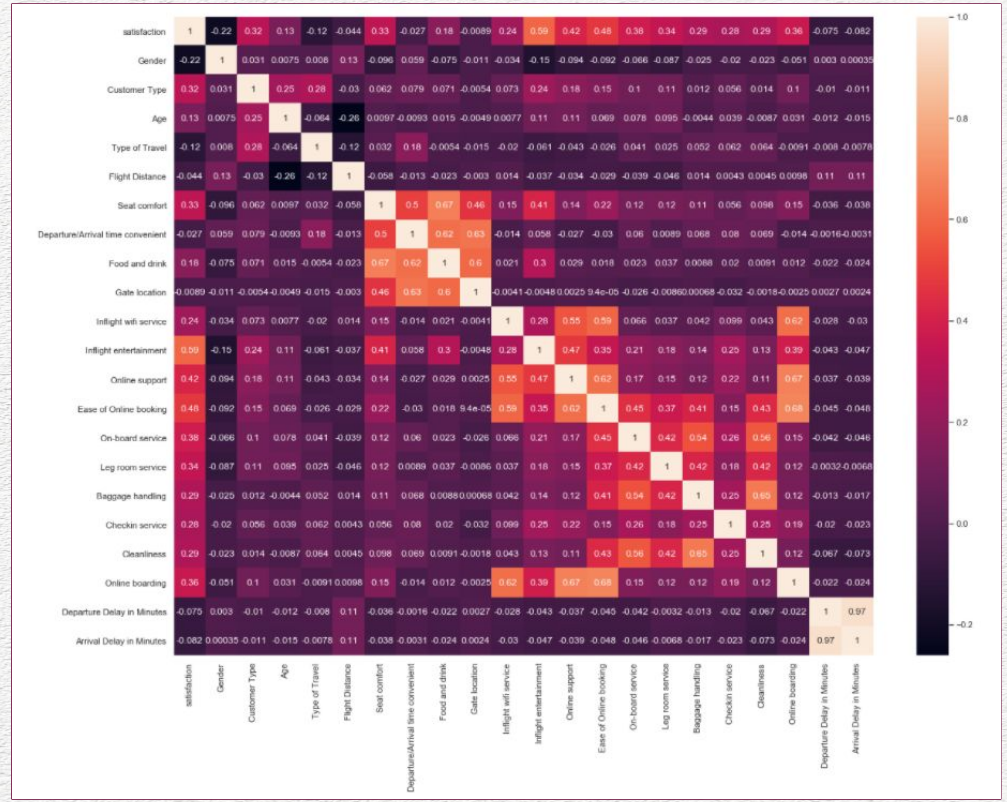
Business Class Passengers are more content with the flight services than the Economy class.



Passengers who are travelling for personal reasons are more dissatisfied over the passengers flying for Business purpose







- Further we use heatmaps to find correlation between the target attribute(satisfaction) and the other columns. We conclude that:

‘Inflight entertainment’, **‘Ease of online booking’**, **’Online Support’**, **‘On-board service’**, **‘Online booking’** are <u>highly positively correlated to the satisfaction of a passenger</u>

Model Selection

- There are various Machine Learning Algorithms that can be used to train a dataset, hence selecting the most suitable algorithm is essential for accurate and optimum results.
- Firstly, we have **split the dataset** into training and testing data, allotting **30% for testing** & **70% for training.**
- Secondly, We have **implemented** the following 3 models and chosen the most optimum one, based on the **Precision, Recall and F1 Score**.

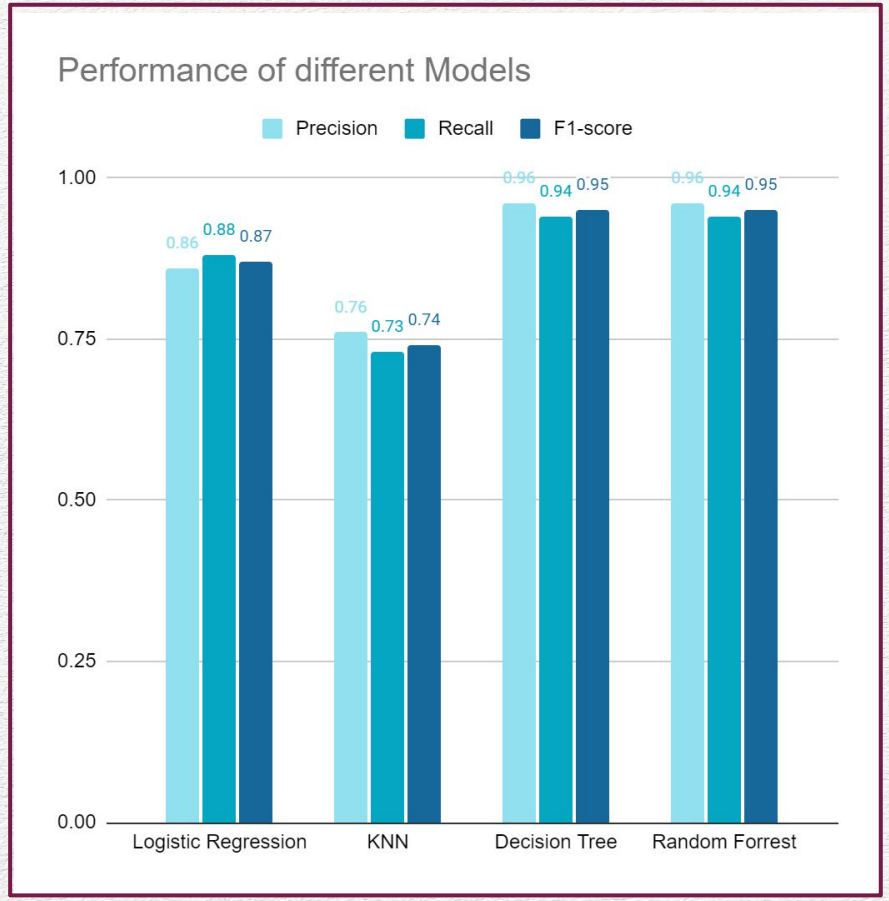
1. Logistic Regression
2. K Nearest Neighbors
3. Decision Tree and Random Forest

The following **parameters** worked best for the respective model classifier.

1. **Logistic Regression** (c=1.0)
2. **KNN** (n_neighbors=10)
3. **Decision Trees**(criterion='entropy', max_depth=30,max_leaf_nodes=1000)
   Random Forest (n_estimators=40, criterion='entropy', max_depth=50,max_leaf_nodes=4100)

The figure shows the performance of the implemented models.
Logistic Regression (0.86) precision , KNN (0.76) precision.

<u>Decision Tree and Random Forests</u> performed the best with the **Precision** of **0.96**

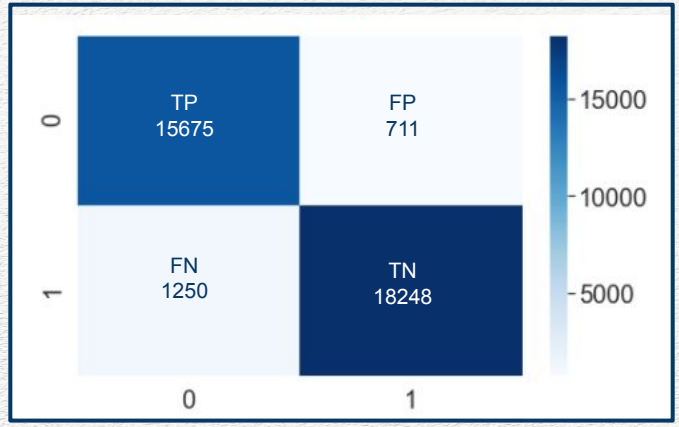


Evaluation

Finally, The evaluation of the model gives us a **Precision of 96.0%** and an **Accuracy of 95.0%**

The Precision of 96.0% means that when the model predicts a passenger to be satisfied, the model is confident that the prediction is 96.0% accurate and true.

Confusion Matrix



Classification Report

```
In [100]: print(classification_report(y_test, prediction))

              precision    recall  f1-score   support

           0       0.93      0.96      0.94     16386
           1       0.96      0.94      0.95     19498

    accuracy                           0.95     35884
   macro avg       0.94      0.95      0.95     35884
weighted avg       0.95      0.95      0.95     35884
```

Passenger Satisfaction Level

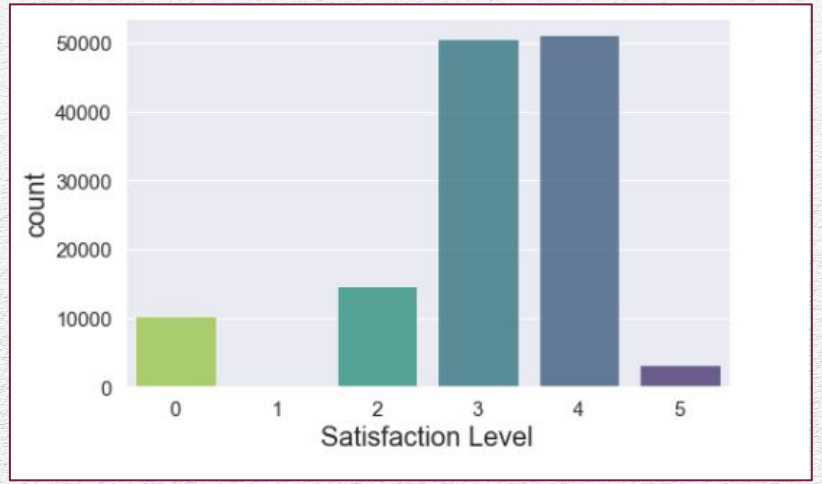
- Now we **calculate the Overall Satisfaction Level** (from 1 to 5, 1 being the lowest and 5 being the highest) for each Passenger.

- Any Passenger with a **0 value** in any of the survey columns **would not be evaluated**, hence such passengers would have the satisfaction level of 0 i.e. **Not Applicable**.

- We calculate the Overall Passenger Satisfaction by taking the **mean of all their 14 survey columns and then rounding it off to its nearest Integer**.

- After evaluating for all passengers, we **append** their respective levels into a new column & add it in the dataframe

Distribution for Overall Passenger Satisfaction Level

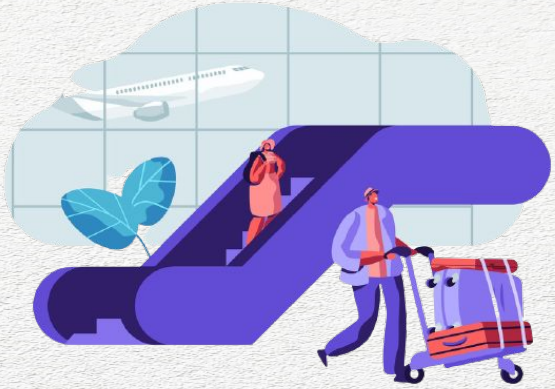


In [121]:
```
data.head()
```

Out[121]:

lass	Flight Distance	Seat comfort	Departure/Arrival time convenient	Food and drink	...	Ease of Online booking	On-board service	Leg room service	Baggage handling	Checkin service	Cleanliness	Online boarding	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction Level
Eco	265	0	0	0	...	3	3	0	3	5	3	2	0	0.0	0
ness	2464	0	0	0	...	3	4	4	4	2	3	2	310	305.0	0
Eco	2138	0	0	0	...	2	3	3	4	4	4	2	0	0.0	0
Eco	623	0	0	0	...	1	1	0	1	4	1	3	0	0.0	0
Eco	354	0	0	0	...	2	2	0	2	4	2	5	0	0.0	0

Business Insights & Key Inferences



1. We found these factors to be affecting the customer satisfaction positively with the following correlation coefficients
   - **Inflight entertainment (0.59)**
   - **Ease of Online booking (0.48)**
   - **Online support (0.42)**
   - **On-board service (0.38)**
   - **Online boarding (0.36)**

Hence, focusing on these factors would result in higher customer satisfaction and as a result higher probability of their return to the airline for travel.

2. We noticed that Airline Online services are highly and positively correlated with satisfaction, hence the company should **invest on a Robust, Easy booking online solution.**

3. We observe that **Male passengers have a higher dissatisfaction** count than that of Female passengers. Figuring out the Male Passenger needs during the flight, and catering to it would help the company tip scales.

4. Passengers travelling for **Personal reasons** have a **higher dissatisfaction** count than that of Travelling for Business.

5. We notice that **major dissatisfied** passengers are from **Economy class**.

6. On studying the Age and Class attributes we get a age group of passengers for each class, such that

- **Major Business class** belongs to the age group of **39-49** years
- **Major Economy class** belongs to the age group of **20-27** years

Since we know maximum dissatisfaction is seen for passengers travelling in economy class, the above insight could be used to **tailor facilities which please the 20-27 age group**, thus increasing Customer Satisfaction.

Thank You !