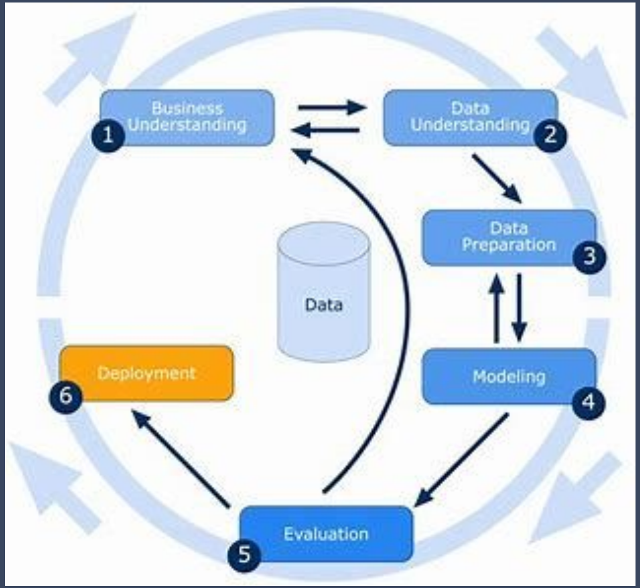
Predicting Hotel Booking Cancellation

Janvier Nshimyumukiza, Ashish Sangai, Sherraina Song, Wenqi (Summer) Zhai

Table of Contents



01

BUSINESS UNDERSTANDING

Business Understanding:

Understanding the Hotel Business

REVENUE

COST

- Guests looking for accommodation
- Food & Beverage
- Entertainment
- Franchise model

- Employee related costs
- Lease/Rent
- Food & Beverage Supply
- Utilities
- Entertainment Supply
- Marketing
- Third-Party Agency Listing Fee

Business Understanding:

Use of Data Analysis

How will the use of Data Analysis improve the hotel’s business?

- **Improve Revenue**
    - Predict cancellations → Improve Revenue by better utilization of resources
    - Predict peak periods → Change cost based on demand
    - Predict customer preferences → Can charge higher for exclusive experiences
- **Decrease Costs**
    - Predict off-season → Lower employee related costs for periods
    - Predict off-season → Optimize F&B costs
- **Improve Customer Experience**
    - Better entertainment → Higher demand for hotel
    - Better entertainment/F&B → Even non-guests can add to revenue

Hotel Business Understanding:

Predict 'Cancellations'

Predicting whether a guest will cancel their booking or not will help the hotel to design marketing strategy to retain the customers that are likely to cancel, and improve the hotel's utilization of its fixed costs and thus, improve the profits.

Objective

Build predictive models to determine whether or not a hotel customer will cancel the booking.

02

DATA UNDERSTANDING

Data Understanding:

Dataset

- Data contains cancellation and guest information from a resort hotel and a city hotel located in the resort region of Algarve and city of Lisbon
- 31 variables describing the **40,060 observations** of resort hotel and 79,330 observations of city hotel.
- Arrivals between the **1st of July of 2015 and the 31st of August 2017**, including bookings that effectively arrived and bookings that were canceled.
- The target variable, **is_cancelled**, is a binary feature with (0,1) as values and thus, we treat this as a classification problem.

Data Understanding:

31 Features

Categorical

- Hotel
- Is_canceled
- Customer_type
- Is_repeated_guest
- Meal
- Country
- Market_segment
- Distribution_channel
- Reserved_room_type
- Assigned_room_type
- Deposit_type
- Agent
- Company
- Reservation_status

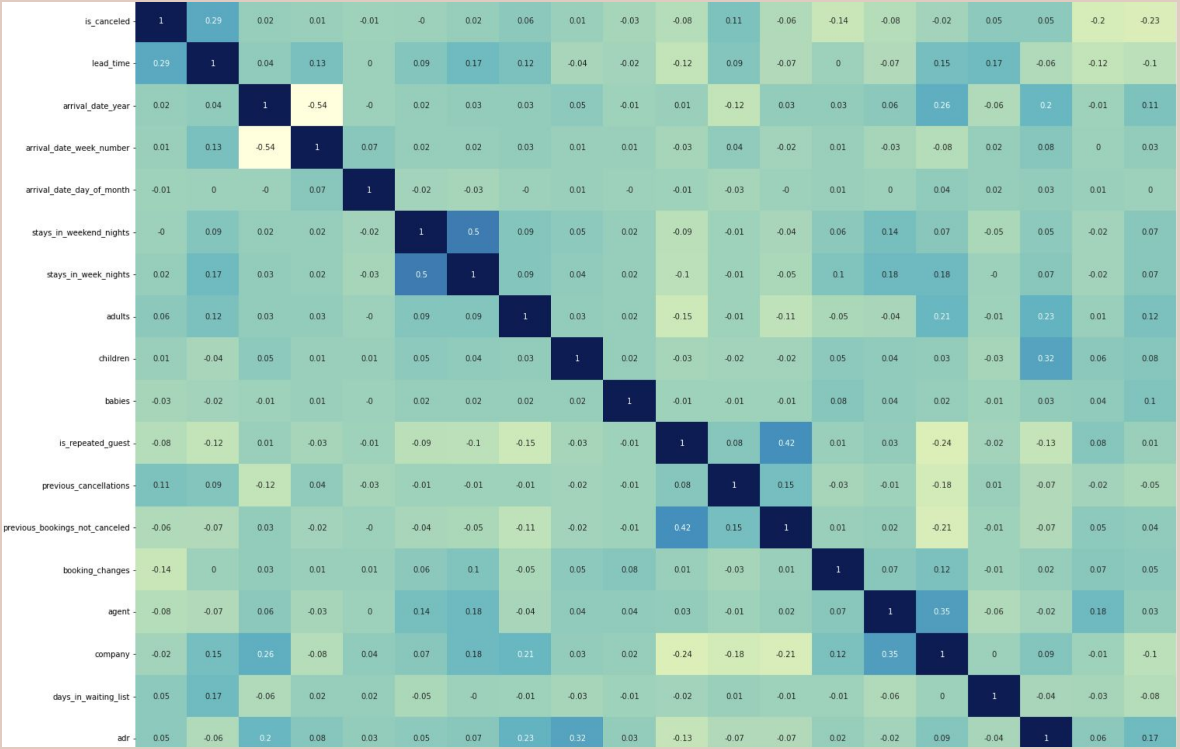
Numerical

- Lead_time
- Stays_in_weekend_nights
- Stays_in_week_nights
- Adults
- Children
- Babies
- Previous_cancellations
- Booking_changes
- Previous_bookings_not_canceled
- Days_in_waiting_list
- Adr
- Required_car_parking_spaces
- Total_of_special_requests
- Arrival_date_year
- Arrival_date_month
- Arrival_date_week_number
- Arrival_date_day_of_month
- Reservation_status_date

Data Understanding:

Exploratory Data Analysis

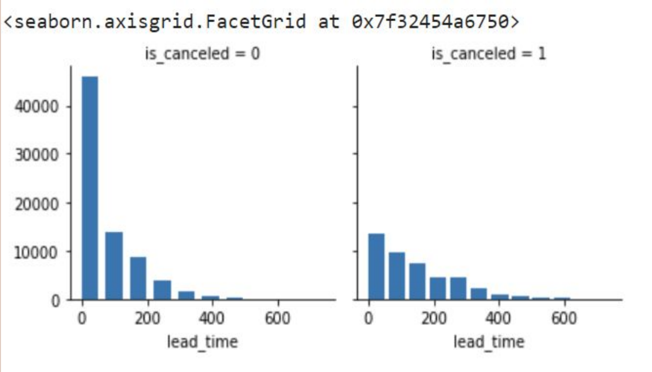
```
corr

is_canceled                       1.000000
lead_time                         0.293123
arrival_date_year                 0.016660
arrival_date_week_number          0.008148
arrival_date_day_of_month        -0.006130
stays_in_weekend_nights          -0.001791
stays_in_week_nights              0.024765
adults                            0.060017
children                          0.005048
babies                           -0.032491
is_repeated_guest                -0.084793
previous_cancellations            0.110133
previous_bookings_not_canceled   -0.057358
booking_changes                  -0.144381
agent                            -0.083114
company                          -0.020642
days_in_waiting_list              0.054186
adr                               0.047557
required_car_parking_spaces      -0.195498
total_of_special_requests        -0.234658
Name: is_canceled, dtype: float64
```



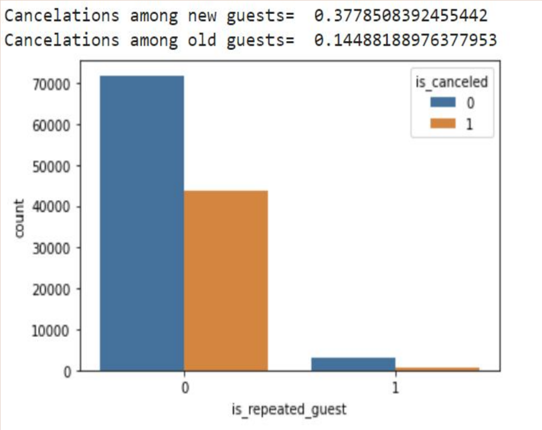
Data Understanding:

Exploratory Data Analysis - Continued

City hotel are more likely to be cancelled than the resort hotel



Guests are less likely to cancel when the booking is made later than earlier



New guests are much more likely to cancel than old guests

Updated Objective:

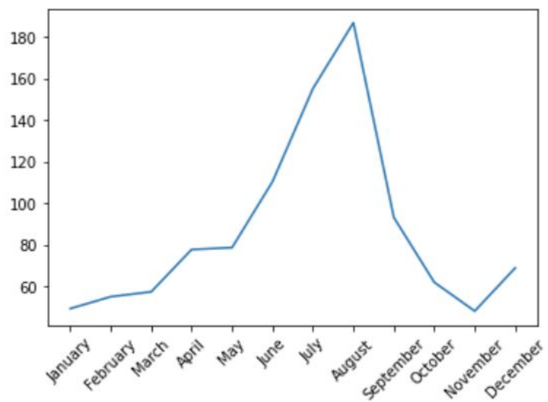
Build predictive models to determine whether or not a RESORT hotel customer will cancel the booking.

Time Trend Analysis:

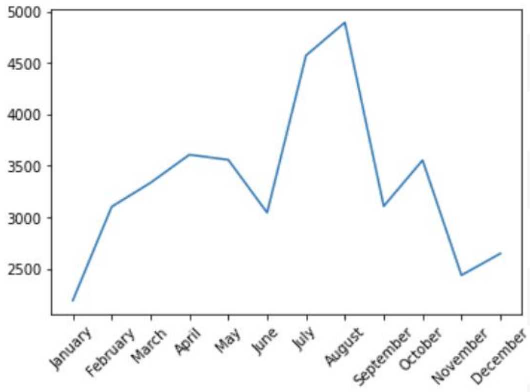
Seasonality in the price per customer & number of guests

month	avg room price
January	49.461883
February	55.171930
March	57.520147
April	77.849496
May	78.758134
June	110.444749
July	155.181299
August	186.790574
September	93.252030
October	62.097617
November	48.273993
December	68.984230

```
plt.plot(final_guests['month'],final_guests['avg room price'])
plt.xticks(rotation=45)
plt.show()
```




```
[33] plt.plot(final_guests['month'],final_guests['no of guests'])
plt.xticks(rotation=45)
plt.show()
```



month	no of guests
January	2193
February	3103
March	3336
April	3609
May	3559
June	3045
July	4573
August	4894
September	3108
October	3555
November	2437
December	2648

We can visualize the seasonality of the booking for resort hotel and get the avg price of 94.95 per person, and we will use this information in the profit curve

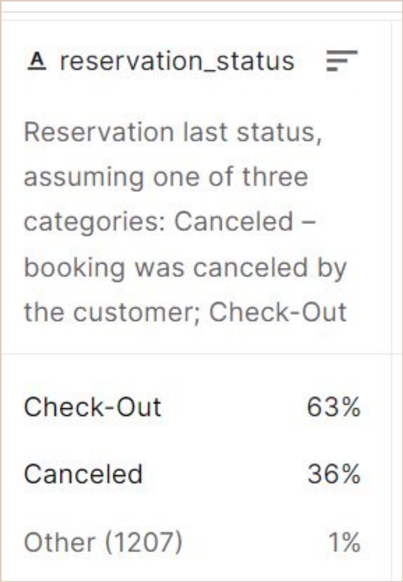
03

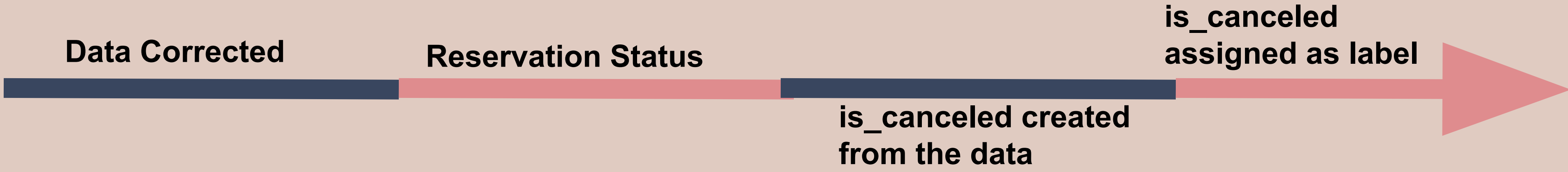
DATA PREPARATION

Data Preparation:

Feature Engineering

- **Data Cleansing** - Missing Values (Fill with NA & Avg)
- **PCA for Dimension Reduction** - kNN only
- **Data Leakage** - reservation status vs. is_canceled





Data Preparation:

Feature Engineering - Continued

- **Apply log transformation to all numerical variables with large variance**
- **Encode Categorical Variables → Binary/Ordinal Numerical Variables**
  - **Binary**: Agent, country
  - **Ordinal**: meal, market_segment,distribution_channel, reserved_room_type, deposit_type,customer_type
- **Drop the variables that are not useful for prediction**
  - Hotel, company, arrival_date_year, reservation_status_date

```
df_RH.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40060 entries, 0 to 40059
Data columns (total 27 columns):
```

04

MODELING

Modeling:

Overview

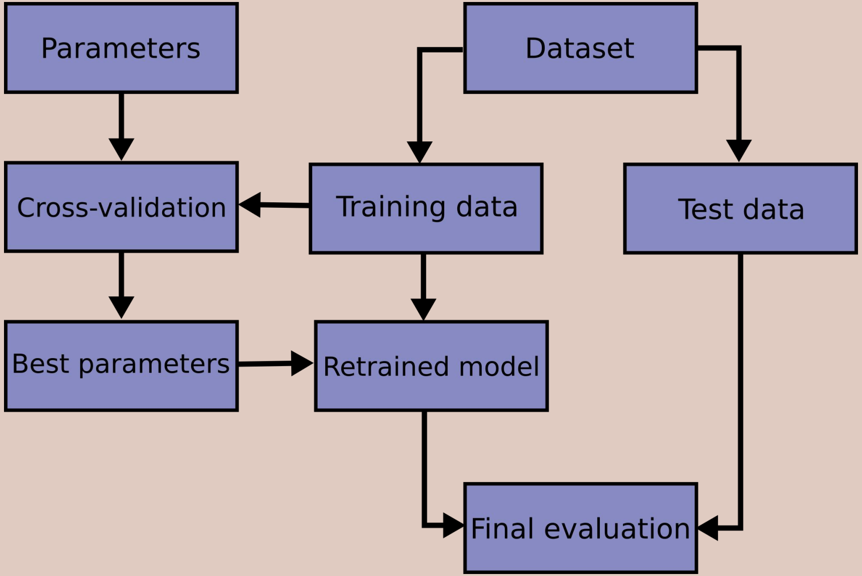
- Tried out 3 models: **Logistic Regression, k-NN, Decision Tree** to determine the best performing one using the following methods:
  - Utilized a grid search for hyperparameter tuning
  - Model-specific feature engineering: dimension reduction using PCA for the k-NN model
    - Run into the issue of curse of dimensionality and thus we reduce the dimensionality by using a subset of features

Modeling:

Grid Search Cross-Validation

Process:

- Split the dataset using train_test_split into 60:40 ratio & in a stratified way.
- Used Grid Search Cross-validation to get the best model using the training set
  - 5 folds
  - We didn’t use nested cross validation as we have a fairly large dataset
- Evaluation: f1 score metric for grid search validation



Modeling:

Grid Search for Hyperparameter Tuning

	Logistic Regression	Decision Tree	kNN
Parameters Grid	{'penalty':['l1', 'l2'], 'solver':['lbfgs', 'liblinear'], 'C':[1e5, 1e4, 1e3, 1e2, 1e1, 1e0, 1e-1, 1e-2, 1e-3]}	'max_depth': range(5,30), 'criterion':['gini','entropy'], 'min_samples_leaf':[2,3,4,5], 'min_samples_split':[2,3,4,5]}	'n_neighbors': [3,5,7,9,13,17,21], 'weights':['uniform','distance'], 'p': [1,2]}
Best Parameters found (Using F1 score)	{'C': 10.0, 'penalty': 'l1', 'solver': 'liblinear'}	{'criterion': 'gini', 'max_depth': 16, 'min_samples_leaf': 2, 'min_samples_split': 5}	{'n_neighbors': 13, 'p': 1, 'weights': 'distance'}

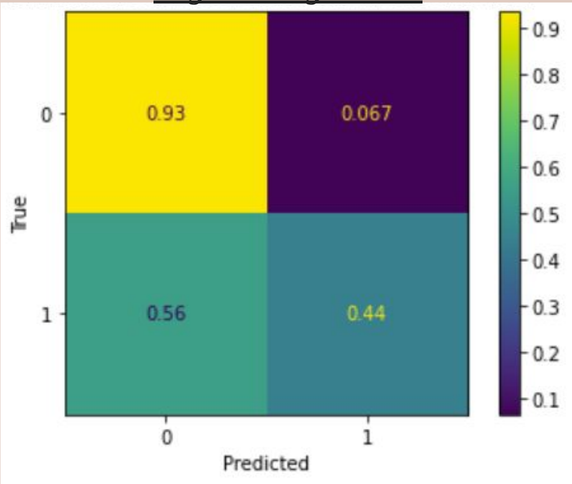
05

EVALUATION

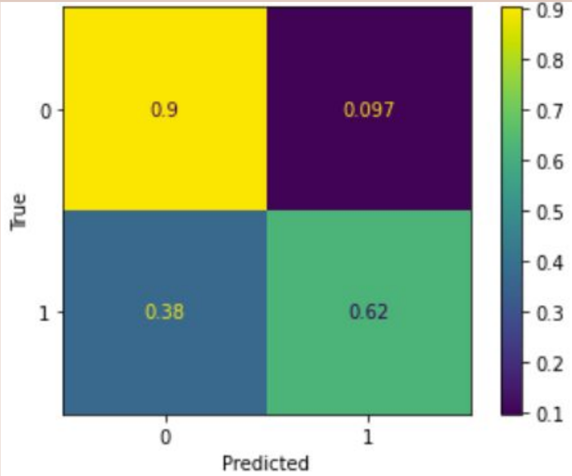
Evaluation:

Comparison (Classification report & Confusion Matrix)

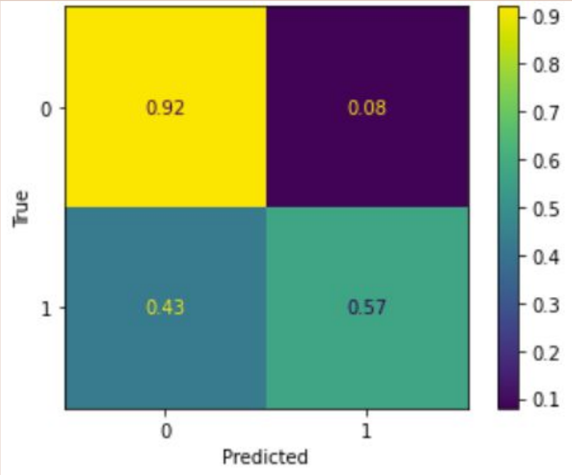
Logistic Regression



Decision Tree



kNN

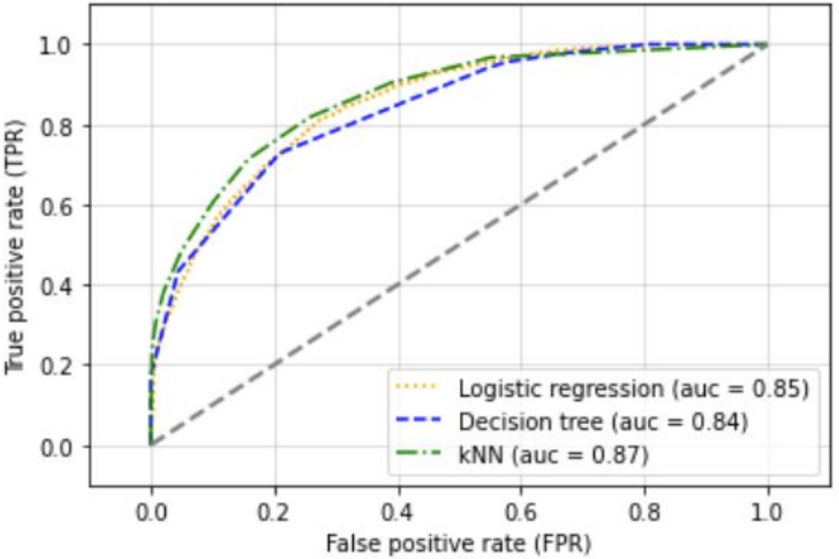


	Logistic Regression	Decision Tree	kNN
F1 score Achieved	0.55	**0.66**	0.65

Evaluation:

ROC AUC

```
10-fold cross validation:

ROC AUC: 0.71 (+/- 0.20) [Logistic regression]
ROC AUC: 0.79 (+/- 0.05) [Decision tree]
ROC AUC: 0.59 (+/- 0.16) [kNN]
```



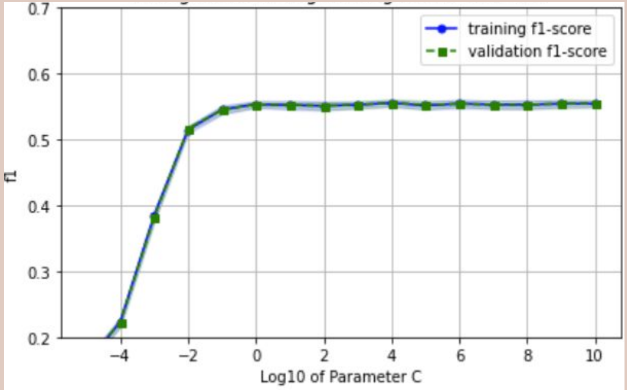
- Decision Tree is the best performing model among the three, with an AUC of ~0.85

- It has a higher true positive rate and a lower false positive rate

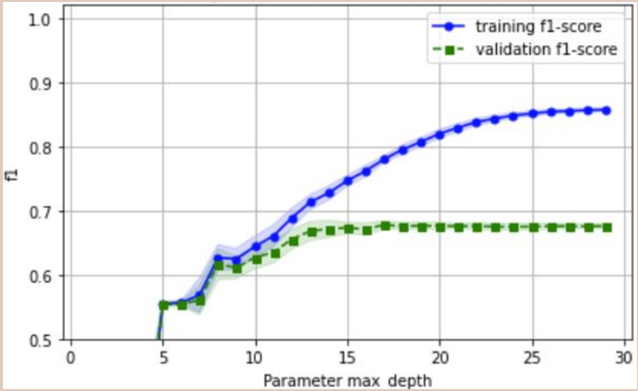
Evaluation:

Fitting Graph

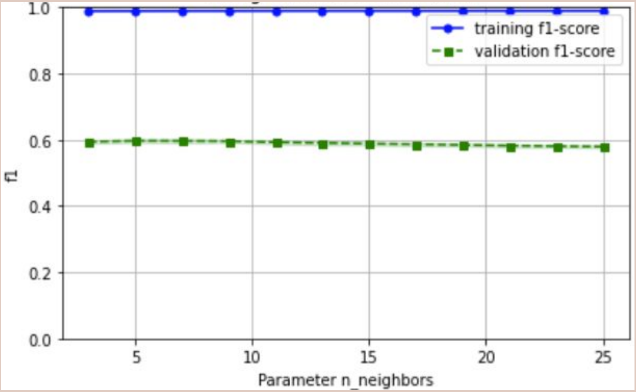
Logistic Regression



Decision Tree



kNN

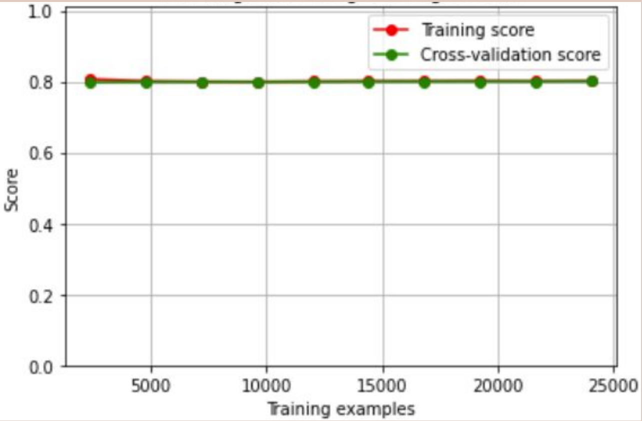


- For Logistic Regression, the best parameter C, we found is 1
- For Decision Tree, the optimal parameter for the depth is 13, and the training set for the tree started to suffer from overfitting when the depth increases
- For kNN, any value for the *k* is not making a big difference

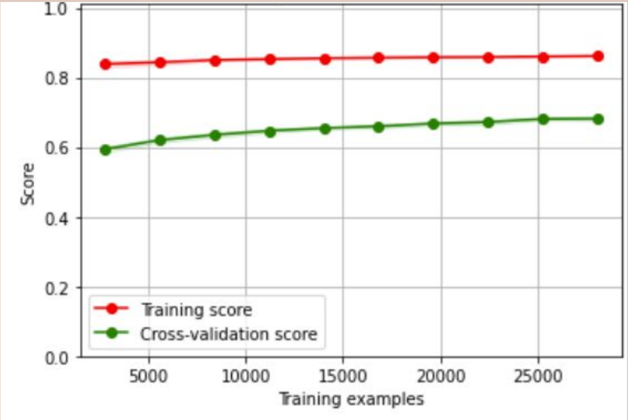
Evaluation:

Learning Curve

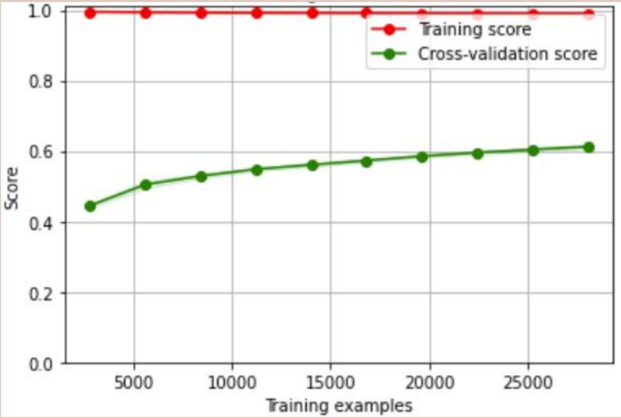
Logistic Regression



Decision Tree



kNN



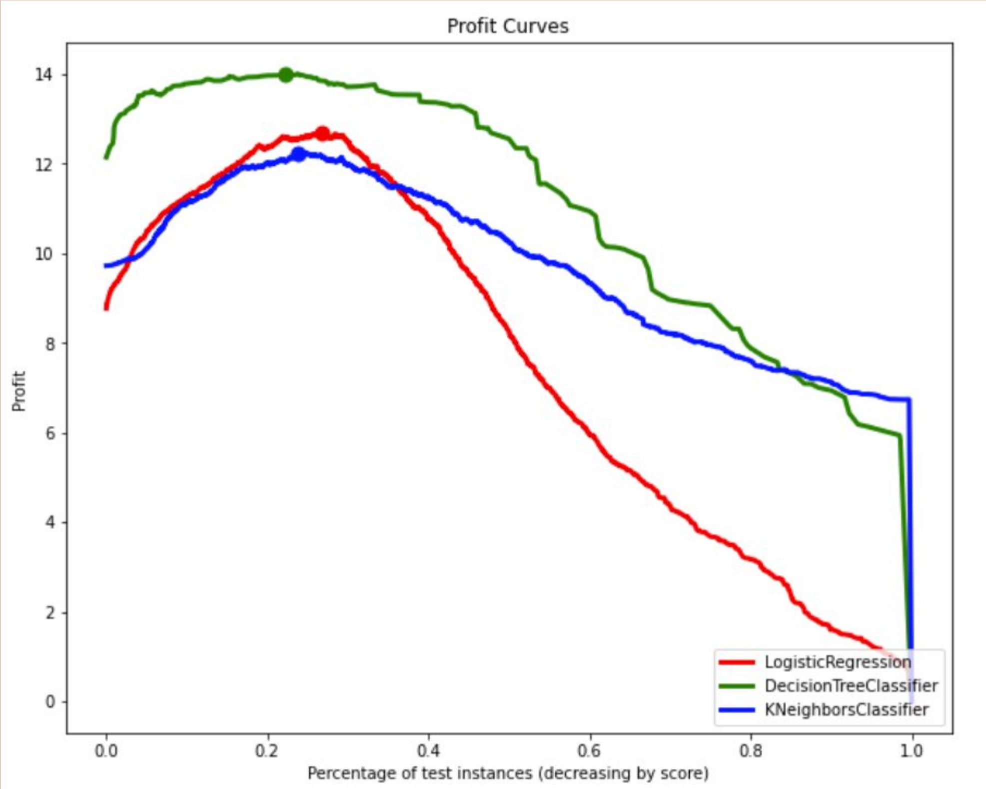
- Increasing the training data size can help improve the performance of kNN and Decision Tree, but not for Logistic Regression (we can't help it:))

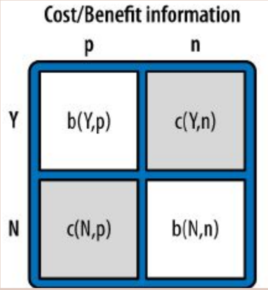
06

DEPLOYMENT

Deployment:

Cost Matrix & Profit Curve





	Actual Cancel	Actual Not Cancel
Pred. Cancel	74 TP	-20 FP
Pred. Not Cancel	0 FN	0 TN

- **Decision Tree** performed the best even when the percentage of customer we can target at changes over time
- Given a budget to target at ~20% of the customers, they will have an expected profit at around **$14** for decision tree

Assumption: Giving a $20 coupon will retain cancelled customers

Deployment:

Implementation

The predictive model can be implemented to determine if a user is likely to cancel the booking or not & thus, retain customers.

This predictive result helps the hotel to:

- Redesign their cancellation policy
- Maximize profit by targeting a specific amount of customers based on the budget
- Reduce cancellation rate by designing a marketing strategy to target customers who are likely to cancel, like offering them coupons

This case study may also help the hotel to:

- Get a better understanding of customer profiling on which group is more likely to cancel
- Given the competitor’s data, evaluate whether the cancelation rate is higher/lower

Thank you
Any questions?