★★★★★

Prediction Of Hotel Booking Cancelation with Classification Models

Kristy Yang

Problem

- ❖ Every hotel in the world faces a same issue in daily operation : potential booking cancellations.
- ❖ Too many cancellations will obviously have negative impact on the hotel's profit and revenue.

Business Opportunity & Solutions

- ★ EDA to find relations between variables and target
- ★ Classification models for prediction
- ★ Provide practical suggestions

Methodology

★★★★★

Collection Data

- Available from kaggle
- 119K observations with 36 features
- Both numerical and categorical
- Target: 0 represents “not canceled; 1 represents “canceled”

Data Cleaning & EDA

- Deal with Missing data, correct data types.
- Drop useless feature
- Converting categorical features to dummy variables
- EDA to show relationship between target and features

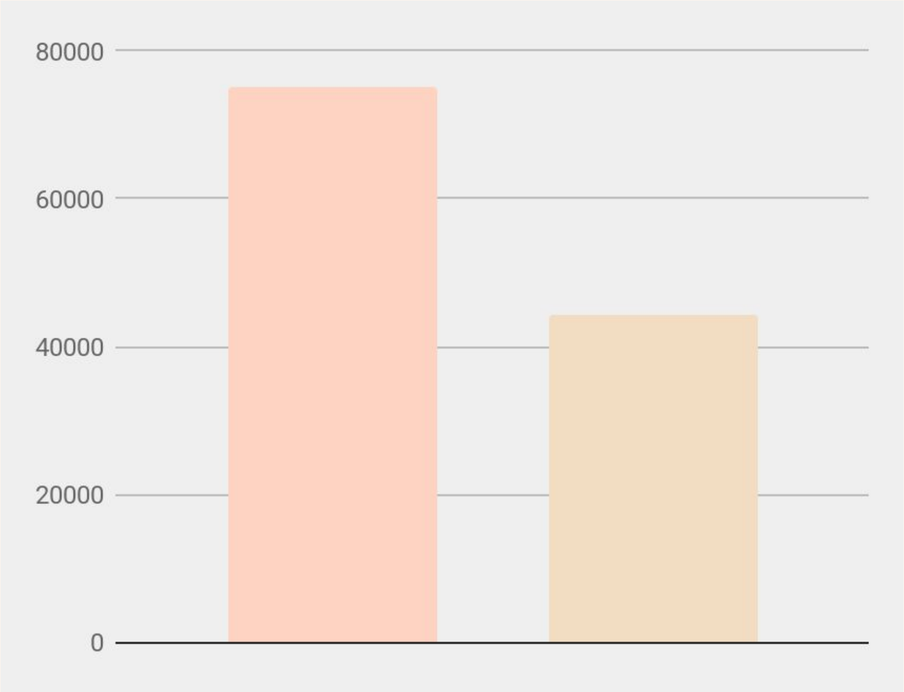
Baseline Models

- Knn
- LogisticRegression
- Decision Trees and Ensembling
    - Decision Tree
    - Random Forest
    - Extra Tree
    - AdaBoost
    - Gradient Boosting
    - Voting Classifier
    - Stacking Classifier
- Bernoulli Naive Bayes
- Gaussian Naive Bayes

Expand and refine model

Tuning and Cross validation for Random Forest Model

Data Overview



1

No_canceled: 75K
Canceled : 44k
Not Imbalanced

2

Useful Features:
18 Numerical features
9 Categorical converted to dummies

Metrics

★★★★★

- ❖ If misclassify a non_canceled booking as canceled:
  - ➢ Potential consequence: overbooking, damaged reputation, losing customers
  - ➢ "Precision" matters

- ❖ If misclassify a canceled booking as non_canceled:
  - ➢ Potential consequence: underbooking, low profit and revenue
  - ➢ "Recall" matters

- ❖ Combined matrics: "F1" score

RandomForest
Wins!

	f1_train	accuracy	precision	recall	f1_test
knn	0.786253	0.778164	0.722949	0.661114	0.690650
logit	0.697836	0.803375	0.819330	0.609459	0.698981
decision_tree	0.989003	0.834157	0.776582	0.782312	0.779436
random_forest	0.988999	0.874864	0.879556	0.771579	0.822037
extra_tree	0.989003	0.866781	0.860413	0.769119	0.812209
adaboost	0.721091	0.818075	0.840640	0.634615	0.723242
gbm	0.808347	0.856102	0.863804	0.731105	0.791934
bernoullinb	0.617582	0.736787	0.683709	0.553220	0.611581
gaussianb	0.609986	0.625681	0.500212	0.793046	0.613475
stackedclassifier	0.984810	0.866530	0.892755	0.731552	0.804154
votingclassifier	0.988981	0.871137	0.867100	0.774709	0.818305

	AUC
logit	0.738471
random_forest	0.852452
extra_tree	0.846546
adaboost	0.781283
gbm	0.831034

Baseline Models

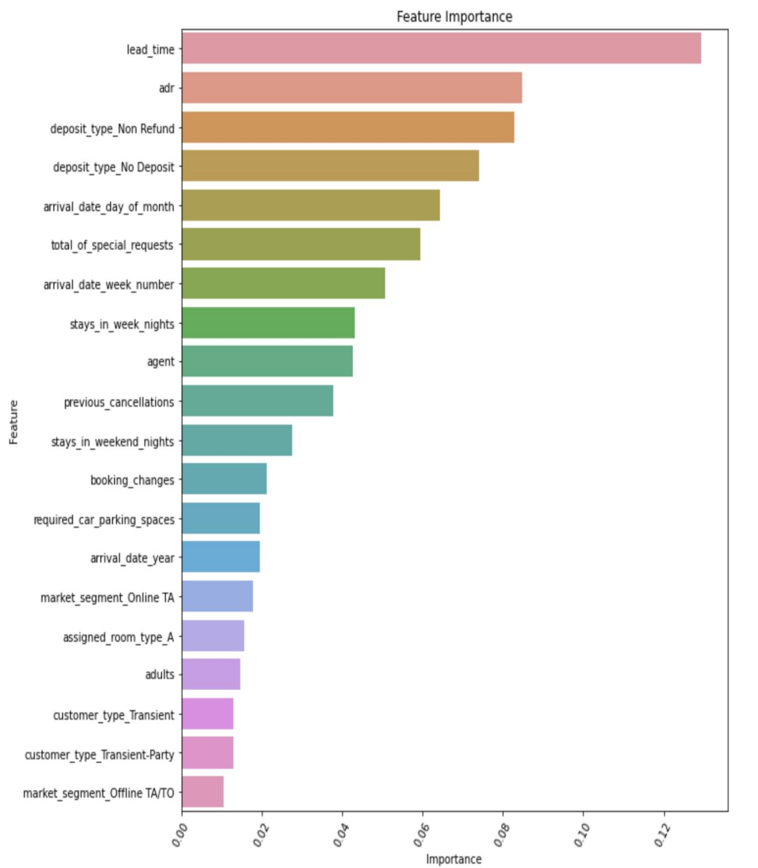
Model Tuning & optimization

- Choose Random Forest
- Adjust hyper parameters
- Results:
-

f1 score of training data is 0.88136:
accuracy score of test data is 0.86515:
precision score of test data is 0.88939:
recall score of test data is 0.73088:
f1 score of test data is 0.80238:
confusion matrix of test data is:
 [[14121   813]
 [2407 6537]]

★★★★★

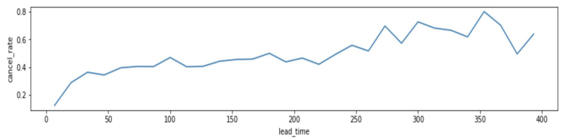
Feature Importance



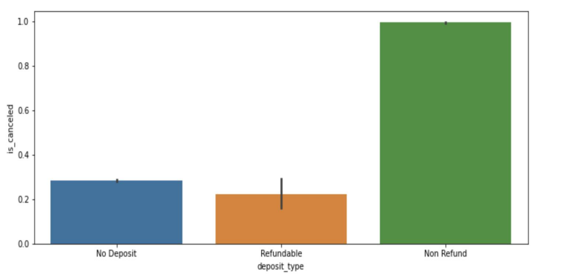
	Feature	Importance
0	lead_time	0.129411
15	adr	0.084846
73	deposit_type_Non Refund	0.082772
72	deposit_type_No Deposit	0.074139
3	arrival_date_day_of_month	0.06444
17	total_of_special_requests	0.0595
2	arrival_date_week_number	0.050688
5	stays_in_week_nights	0.043152
13	agent	0.042669
10	previous_cancellations	0.03779
4	stays_in_weekend_nights	0.027517
12	booking_changes	0.021236
16	required_car_parking_spaces	0.019505
1	arrival_date_year	0.019504
43	market_segment_Online TA	0.017707
60	assigned_room_type_A	0.015633
6	adults	0.014587
77	customer_type_Transient	0.013066
78	customer_type_Transient-Party	0.012901
42	market_segment_Offline TA/TO	0.010494

Lead_time

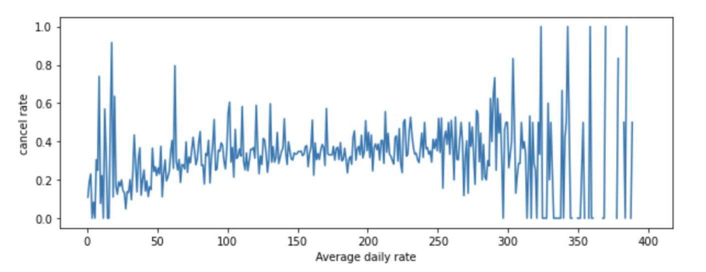
Number of days that elapsed between the entering date of the booking into the PMS and the arrival date



Deposit Type

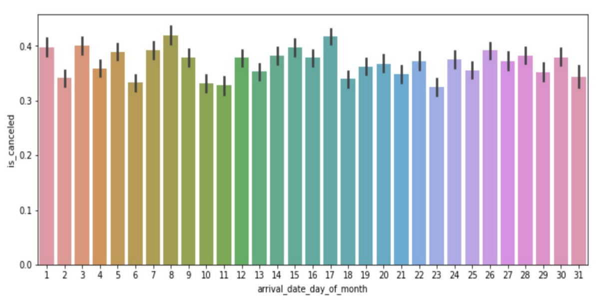


adr: Average daily rate

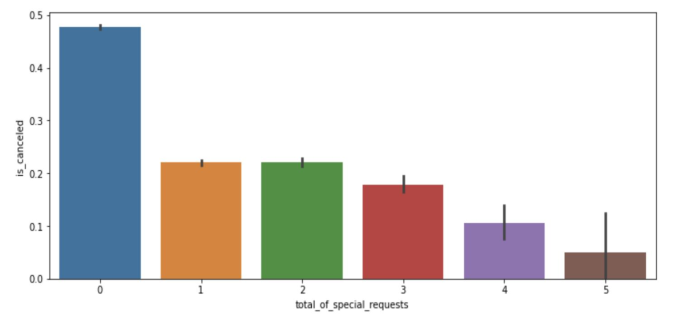


Model Insights - Most important features

Arrival date day of Month

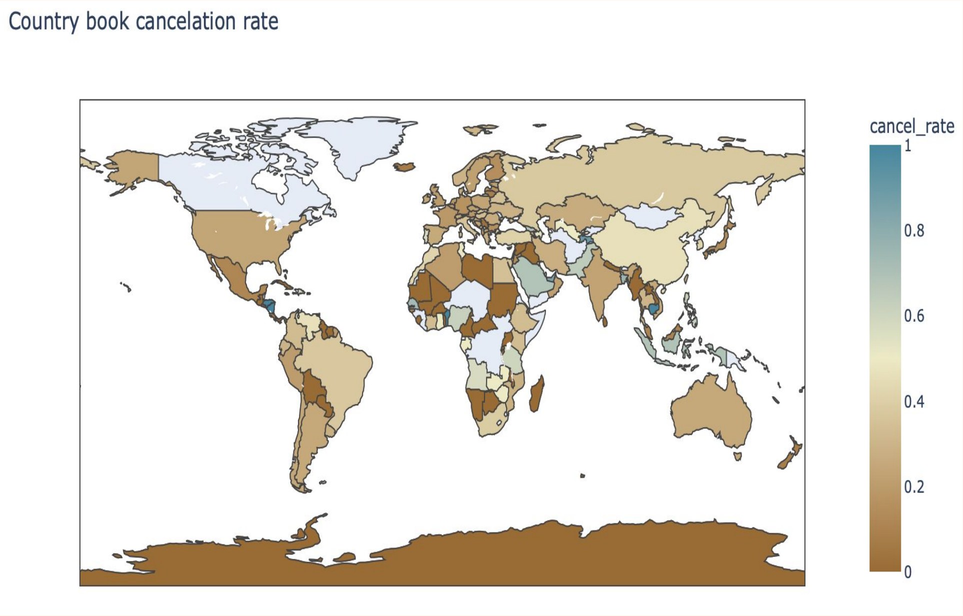


Total special requests



Model Insights - Most important features

Location insight



- 100+ countries
- Drop the feature to simply model and decrease model training time
- Can be used as inference

Suggestions

I	Promote	❖ Deposit refundable booking; ❖ Market in low cancelation countries/segment
II	Adjust	Average daily rate to a reasonable range
III	Provide	Service that meets special requests of guests
IIII	Collect	booking info for each arrival date, perform prediction with model and calculate cancellation probability. Adopt measurements to decrease cancellation rate.

Thanks

& Questions?

CREDITS: This presentation template was created by **Slidesgo,** including icons by **Flaticon**, infographics & images by **Freepik**