Luxury

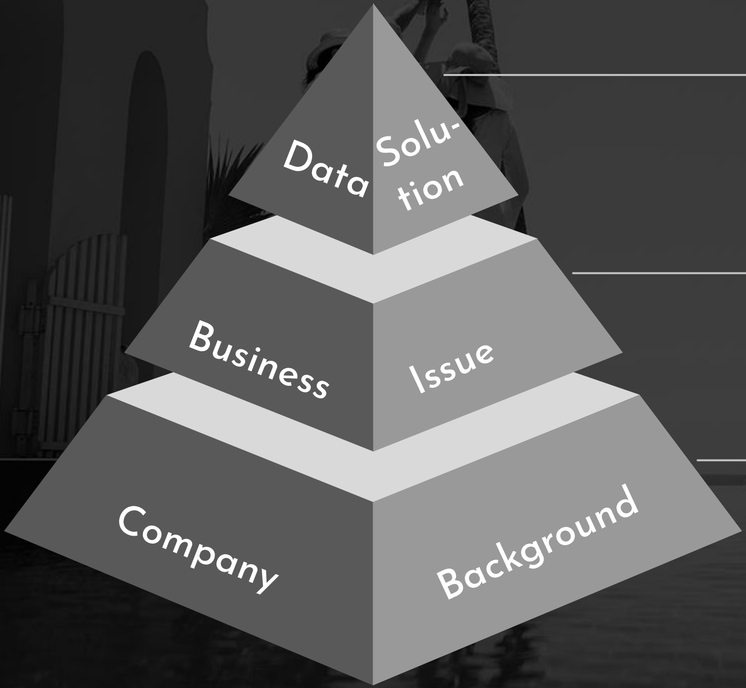
Hotel Booking Cancellation Prediction

Team 4

Xiran Li, Jane Liu, Kelsey Liu,
Haiyue Wang, Kairan Zhong, Yue Zhang

Table of Content

Business Challenge



3 Employ data mining techniques to

- unravel underlying patterns of cancellation
- predict if each room reservation will be canceled

2 "Unpredictable" room reservation cancellation

- deteriorates customer satisfaction
- harms operational efficiency

1 Two Portuguese hotels (1 city, 1 resort) attempt to

- increase revenue
- improve cost-effectiveness

Reference: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand?datasetId=511638&sortBy=voteCount

Exploratory Analysis

26 Months
JUL 2015 - AUG 2017

Lisbon, Portugal
Travel season Sep, Oct

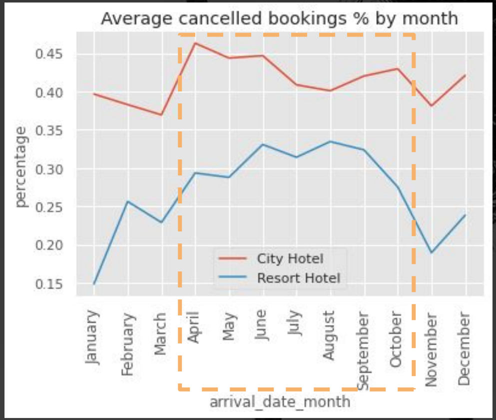
City : Resort
61.5% : 38.5%

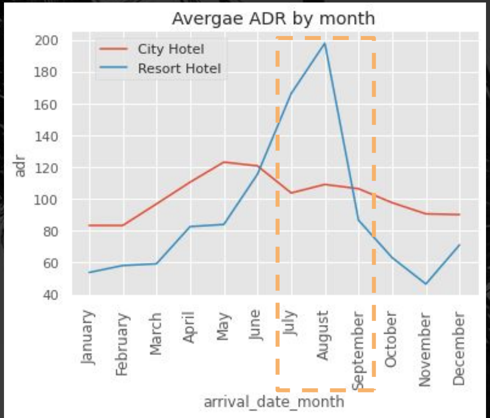
Algarve, Portugal
Travel season Jun, Jul, Aug

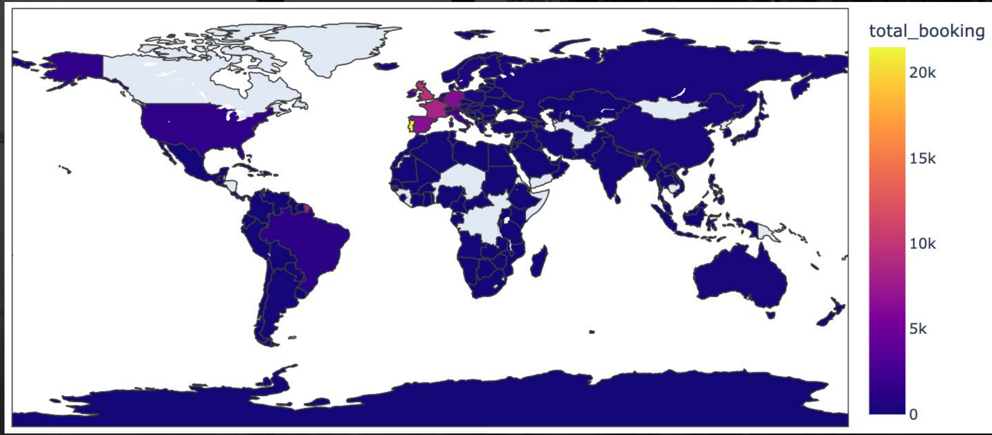
1) Average Cancelled %:
Overall, both hotel types show quite similar fluctuation patterns, while the cancelled bookings percentage for **City Hotel** is consistently higher.

2) Average ADR:
There is huge seasonality effect (best time to visit) for **Resort Hotel** while the ADR fluctuation in **City Hotel** is comparably smaller.

3) Traveler demographic (geo):
Most bookings are from **European** countries, among which PRT (Portugal) makes up the biggest share (18%).







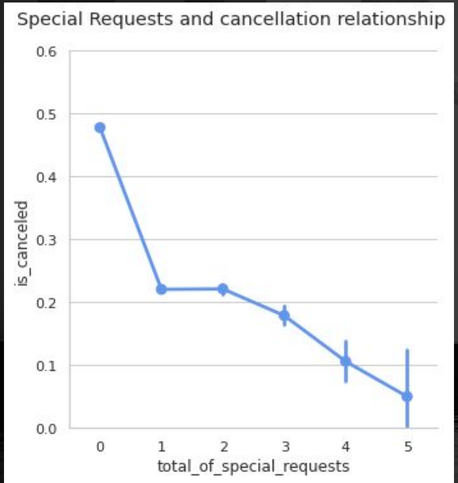
Exploratory Analysis

119k observation
30 original features
1 response (is_cancled)

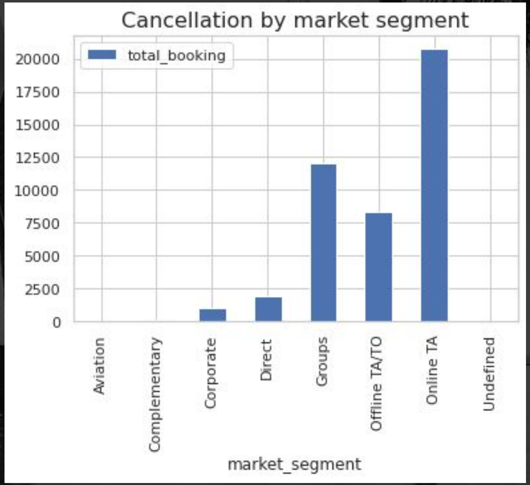
Missing Value
< 1%: Mode Imputation
~15%: New Label
~95%: Drop Column

40%

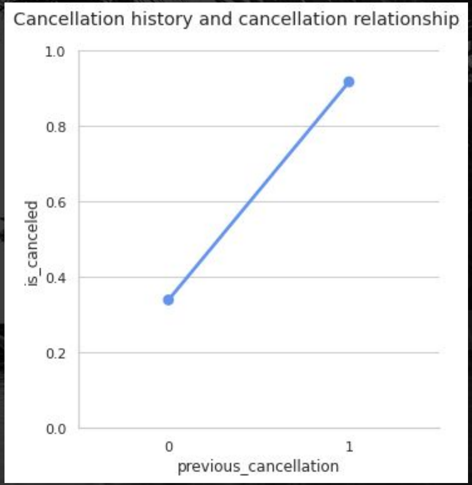
canceled : not cancelled = **37: 63**
No imbalanced issue



The **more** requests, the **less likely** a traveler will cancel.



Among cancelled bookings, **online travel agencies** (OTAs) accounts for the largest proportion.



Travelers with cancellation records are **more likely** to cancel.

Feature Transformation & Engineering

1. Drop some features

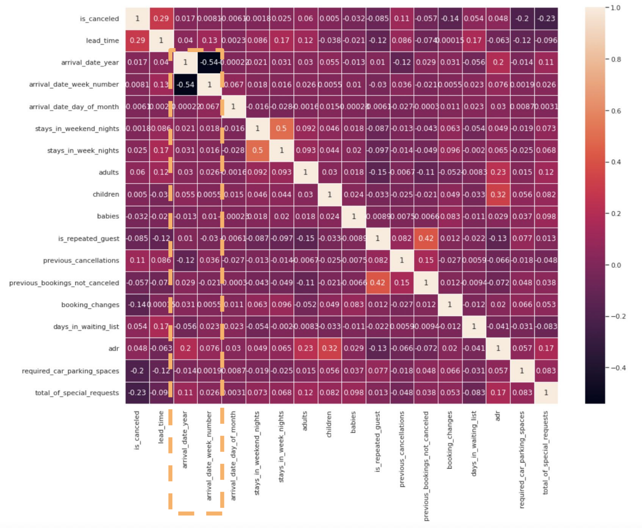
- **Useless features**
  1) meaningless features, eg., agentID
  2) some text columns that cannot be encoded
  ....
- **Features that will change over time**
  e.g., “booking_changes”
- **Highly- correlated pairs**
  save the more informative one of each highly-correlated pair.
  E.g.:
  Save:“arrival_date_week_num”
  Drop: “arrival_date_year

2. Create some new features

- 1) adult + children -> is_family
  2) adult + children + babies -> total_customer
  …

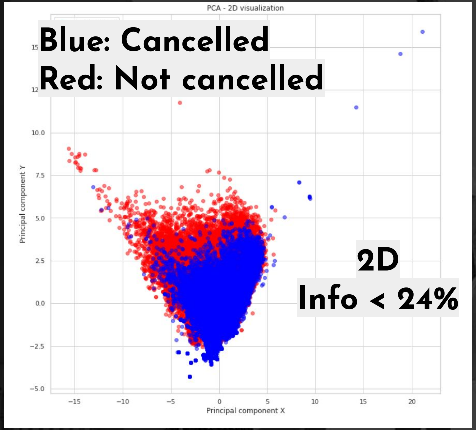
3. Feature Transformation

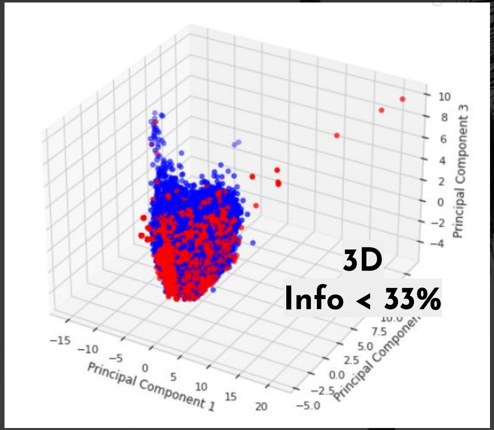
- **For categorical features:**
  1) Manually Encoding
  2) Ordinal Encoding
- **For numerical features:**
  Log transformation for those with large variance

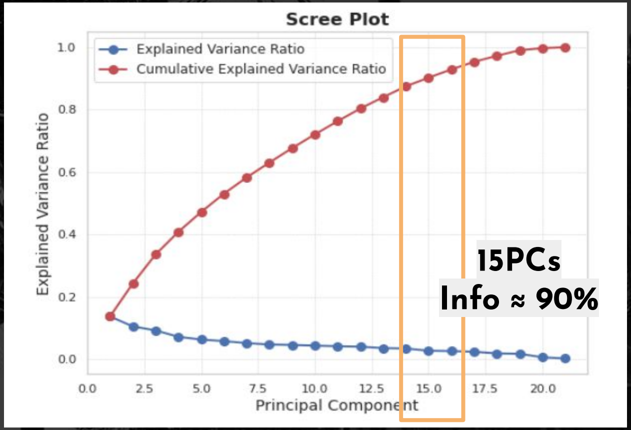


Dimension Reduction

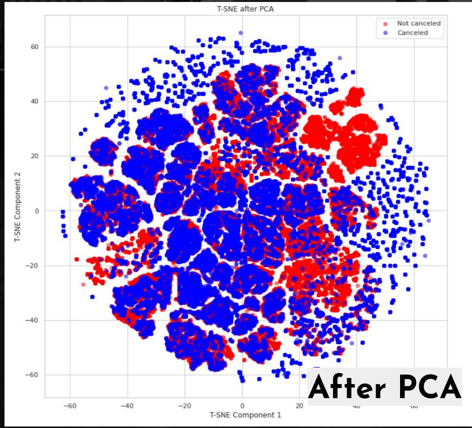
PCA

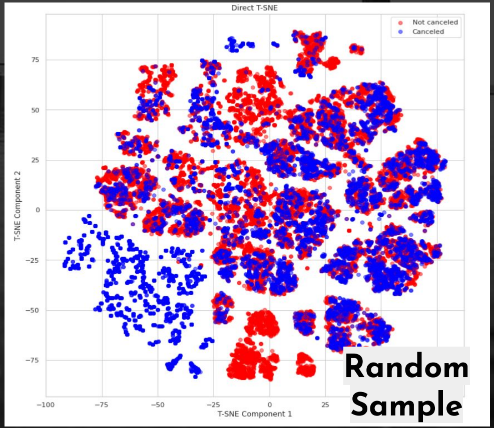






t-SNE





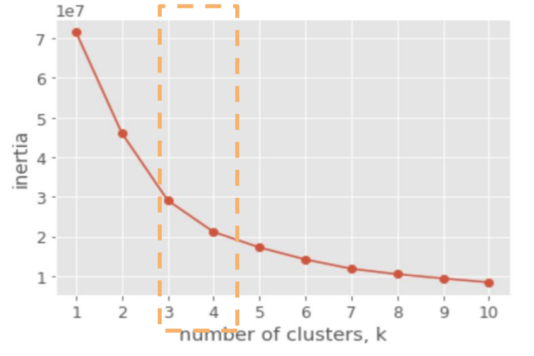
Conclusion:

1. Save the first 15 principal components as a new dataset to do further clustering

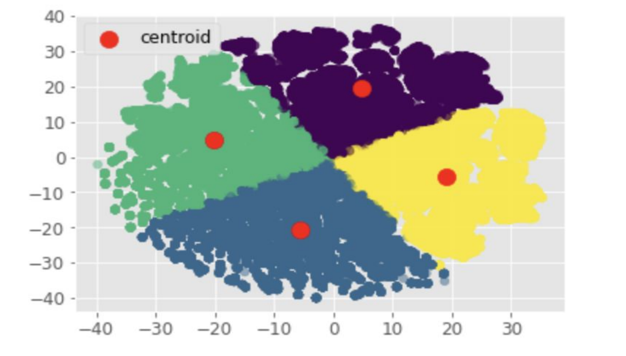
2. More exploration is needed to determine the underlying data patterns and structure

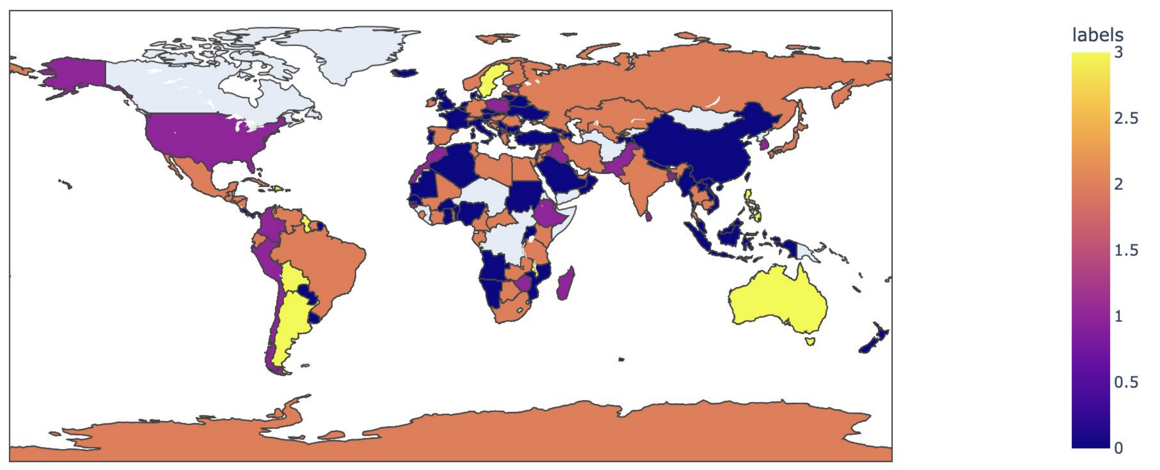
Clustering

Elbow Method



K-Means





Business Traveller (Label 0): More likely to live in a city hotel

Self-driving Lover (Label 1): Needed car parking space

Group Travel Enthusiasts (Label 2): People love to travel in group

Loyal Traveller (Label 3): More likely to be a repeated guest

Classification

	Name	Time	Train_Accuracy	Test_Accuracy	CV_Accuracy	Train_Precision	Test_Precision	Train_Recall	Test_Recall	Train_F1	Test_F1
1	RandomForestClassifier	16.000975	0.995464	0.879941	0.87475	0.996373	0.868217	0.991319	0.799569	0.99384	0.83248
3	XGBClassifier	12.372164	0.882213	0.858658	0.858078	0.867409	0.835357	0.803717	0.77364	0.834349	0.803315
0	DecisionTreeClassifier	2.117703	0.995464	0.834711	0.82849	0.997633	0.772865	0.99006	0.788793	0.993832	0.780748
2	GradientBoostingClassifier	13.723186	0.829495	0.827198	0.827523	0.795211	0.794285	0.724639	0.724475	0.758287	0.757775
6	KNeighborsClassifier	0.014567	0.880919	0.822147	0.819846	0.858497	0.778535	0.811036	0.731344	0.834092	0.754202
4	LogisticRegression	0.903223	0.765746	0.764756	0.769465	0.713145	0.713563	0.611111	0.617255	0.658197	0.661924
5	SGDClassifier	18.601356	0.745706	0.741187	0.746661	0.746919	0.746104	0.470384	0.464305	0.577241	0.572401

Final Model:
XGBoost

Accuracy: 0.86
Precision: 0.84
Recall: 0.77

Important Features:
Required Parking Spaces,
Previous Canceled,
Market Segment...

Insights & Recommendations

Booking Cancellation Challenge

Insights

Important Features:

- Required Parking Spaces
- Previous Canceled
- Market Segment

Customized modeling:

- Seasonal : High vs Low season
- Hotel-specific : City vs Resort

Advices

Based on important features:

- Parking: clear guidance
- Tighten: cancellation policies
- Channel: direct booking

For sustainable improvements:

- Incentives: for non-cancellation
- Marketing: cohort targeting
- Dis-association: OTAs

Thanks