Big Data Analytics of Hotel Bookings

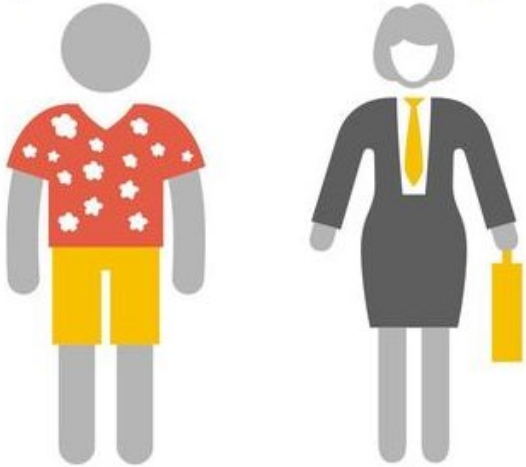
Shuying Yu, Philip Yoon, Fei Xu, Kenneth Liu

Fall 2020

PSTAT 135/235

Stays on weekdays
vs.
Stays on weekends

Booking hotel
5 days head
vs.
737 days ahead

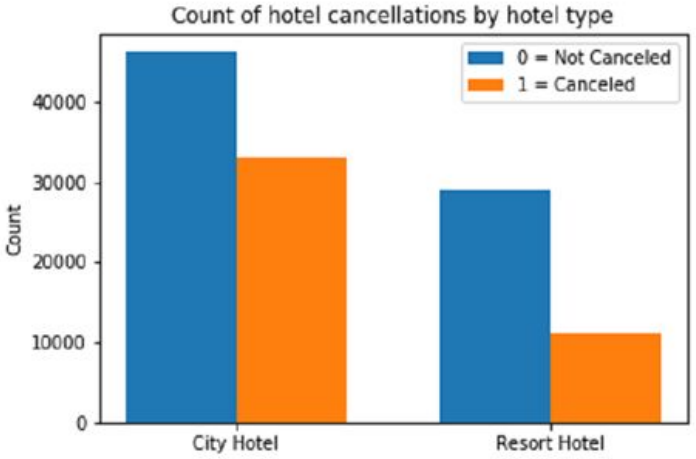


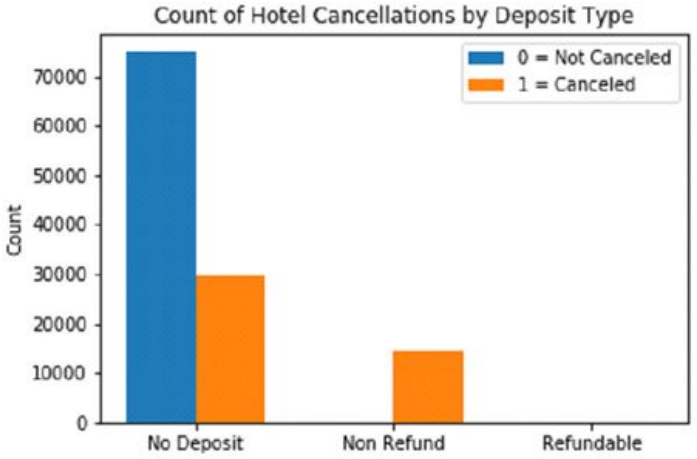
Number of adults
vs.
Number of children

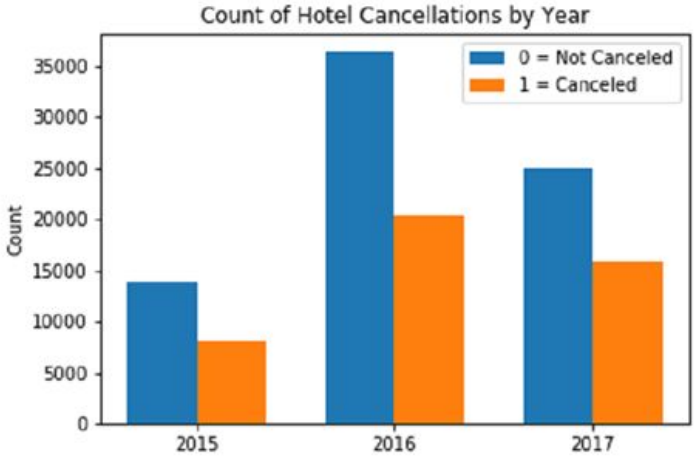
Research Question

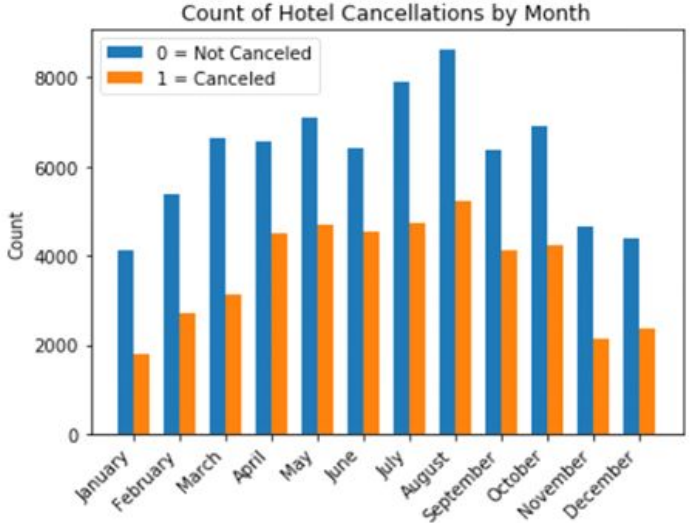
“Can we predict whether hotel bookings will be canceled or not?”

We determined that we are able to predict the status of hotel bookings with high accuracy using *logistic regression* classification


Count of hotel cancellations by hotel type
0 = Not Canceled
1 = Canceled
Count
0
10000
20000
30000
40000
City Hotel
Resort Hotel

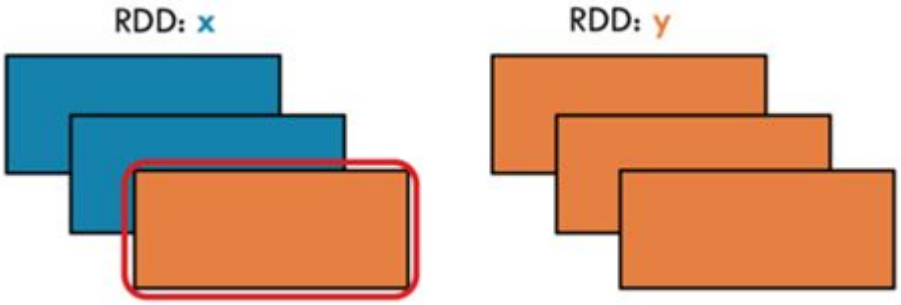

Count of Hotel Cancellations by Deposit Type
0 = Not Canceled
1 = Canceled
Count
0
10000
20000
30000
40000
50000
60000
70000
No Deposit
Non Refund
Refundable

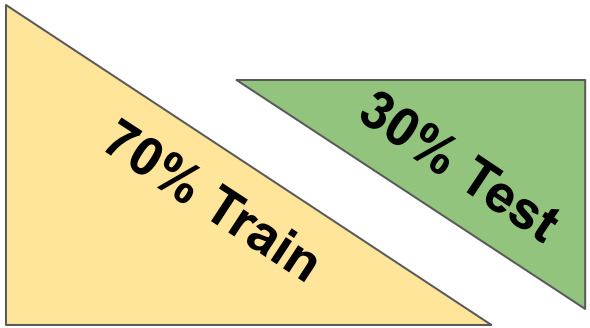

Count of Hotel Cancellations by Year
0 = Not Canceled
1 = Canceled
Count
0
5000
10000
15000
20000
25000
30000
35000
2015
2016
2017


Count of Hotel Cancellations by Month
0 = Not Canceled
1 = Canceled
Count
0
2000
4000
6000
8000
January
February
March
April
May
June
July
August
September
October
November
December

Variable Transformations and Preprocessing

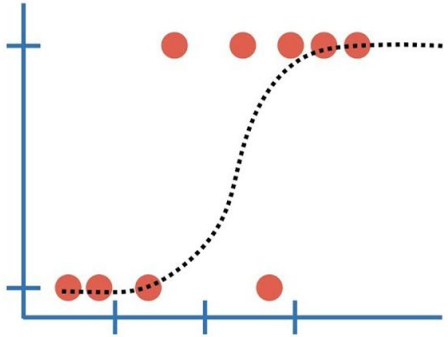
- Remove “NULL” and “N/A”
- Remove empty and redundant variables
- Normalize continuous variables
- Map to labeled point object
- Randomly split data for training and testing

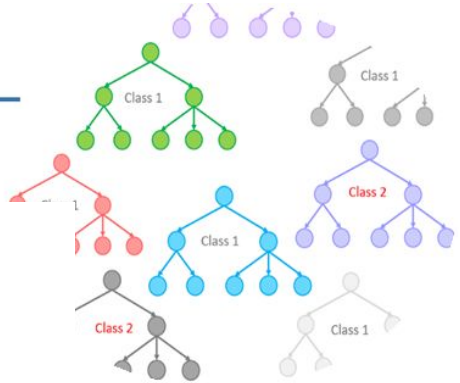


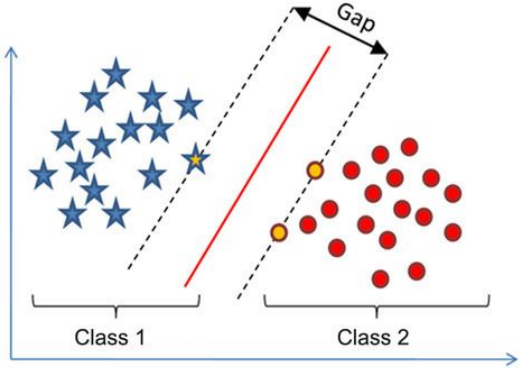


Models Constructed with Tuning Hyperparameters

- **Logistic Regression**
  - Intercept (False, True)
  - Iterations (10, 50, 200)
- **Random Forest**
  - Number of Trees (5, 50, 100)
  - Maximum Depth (5, 15, 30)
- **Support Vector Machine**
  - Intercept (False, True)
  - Iterations (100, 200)
  - Type of Regularizer (l2,l1)







Logistic Regression Model Performance

- Best Logistic Regression model is with no intercept and 100 iterations
- Model is indifferent to adding intercept term
- Not sensitive to iterations

Table 3.2: Model evaluation for logistic regression.

	Accuracy	FPR	FNR	Precision	Recall	F_1	AUC
Base model: no intercept, iteration=100	.8128	.193	.171	.624	.829	.712	.818
Iteration=100, with intercept	.8131	.194	.169	.622	.831	.712	.819
Iteration=10	.7998	.195	.213	.630	.787	.700	.796
Iteration=10, with intercpt	.7993	.196	.213	.628	.788	.699	.796
Iteration=200, no intercpt	.8129	.194	.170	.623	.830	.712	.818
Iteration=200, with intercpt	.8128	.194	.170	.623	.830	.712	.818

Random Forest Model Performance

Table 3.3: Model evaluations for Random Forest.

	Accuracy	FPR	FNR	Precision	Recall	F_1	AUC
Base model: No of Trees=5, max depth=5	.558	.303	.160	.613	.839	.708	.768
Max. Depth=30	.611	.192	.169	.801	.830	.815	.819
No. of Trees=50, Max. Depth=15	.609	.219	.137	.751	.862	.803	.822
No. of Trees=100	.574	.296	.103	.608	.897	.725	.801

- Increasing maxDepth increases accuracy and decreases FPR
- Increasing numTrees increases accuracy slightly and decreases FNR
- Model with maxDepth = 30 is the best model for Random Forest with the highest accuracy and highest precision, which are what we want for this problem

Support Vector Machine Model Performance

Table 3.4: Model evaluation for Support Vector Machine.

	Accuracy	FPR	FNR	Precision	Recall	F_1	AUC
Base model: iteration=100, no intercept	.548	.164	.555	.883	.445	.592	.640
Iteration=100, with intercept	.638	.364	.257	.034	.744	.066	.690
Iteration=200, no intercept	.554	.168	.552	.875	.448	.593	.640
Iteration=200, with intercept	.633	.368	.274	.015	.726	.029	.679
Iteration=200, with intercept regType='l1'	.634	.367	.236	.019	.764	.038	.699
Iteration=100, with intercept regType='l1'	.638	.364	.255	.034	.745	.066	.690

- Adding intercept improves the accuracy but decreases precision.
- Increasing iteration doesn't increase accuracy.
- Precision and F1 score shows that intercept should not be used for this model.
- Use less iteration for saving space

Comparison of models' best performance

Table 3.5: Models with highest accuracy.

	Accuracy	FPR	FNR	Precision	Recall	F_1	AUC
Logistic Regression: iter=100, w/intercept	.813	.194	.169	.622	.831	.712	.819
RF: maxDepth=30	.611	.192	.169	.801	.830	.815	.819
SVM: iteration=100, with intercept regType='l1'	.638	.364	.255	.034	.745	.066	.690

The best design is logistic regression with 100 iterations and intercept.

This model has a high accuracy and a decent precision.

Conclusions and Future Research

- **Best performing model**
  - Logistic regression with intercept
- **Limitations**
  - Latent variables
  - Tools: server has limited memory
- **Future Directions**
  - Study the weights of different variables
  - Performance on more data
  - Tune other parameters of model
  - More visualizations

Thank you!