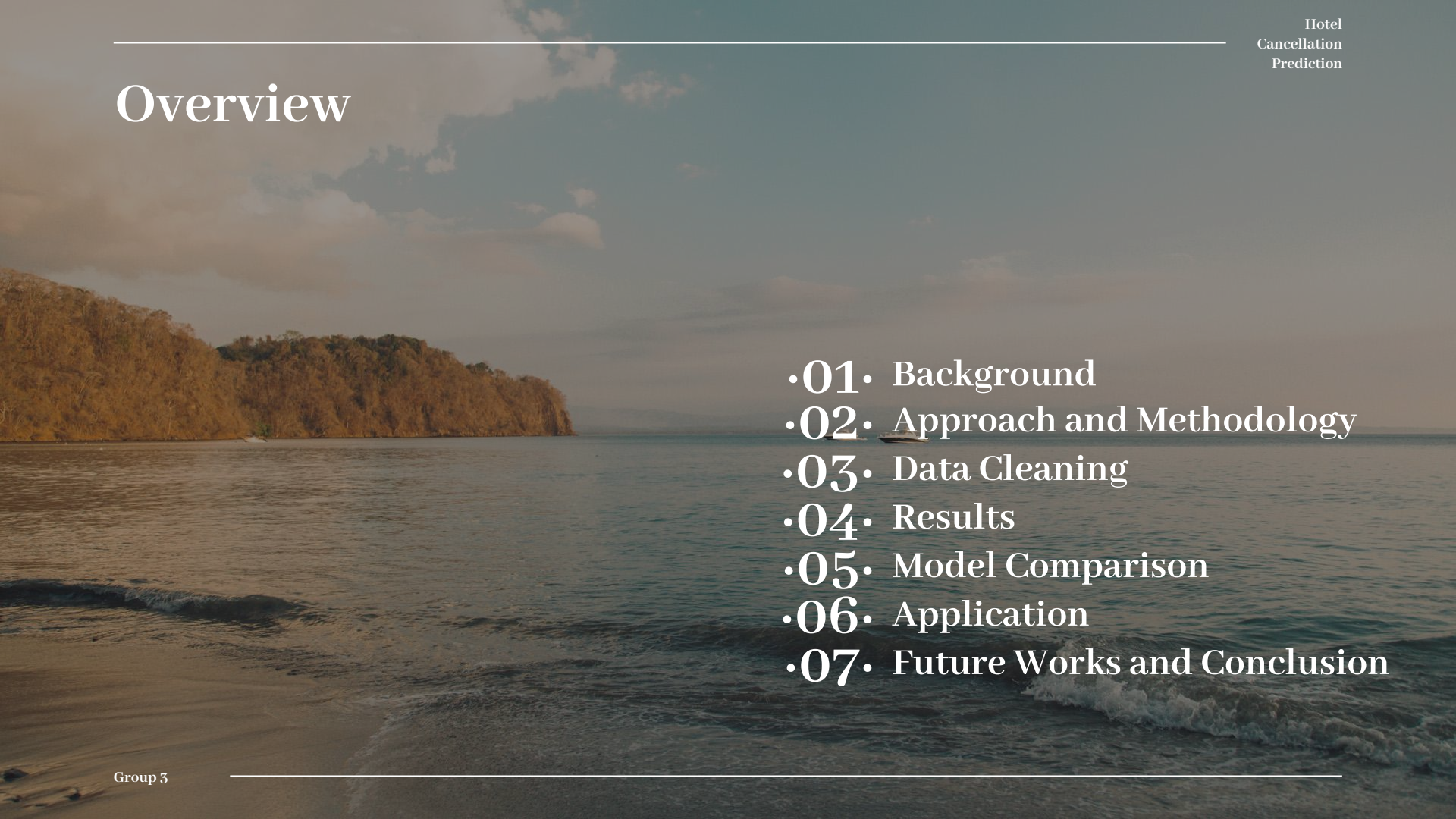
Hotel Cancellation Prediction

Group 39
Josh Bielenberg
Joseph Kim
Nigel Yee
Ying-Kai Huang

Overview



·01· Background
·02· Approach and Methodology
·03· Data Cleaning
·04· Results
·05· Model Comparison
·06· Application
·07· Future Works and Conclusion

·01· Background

Hotel Booking Cancellations

Rooms that are **booked but not filled** / paid for equate directly to **lost potential revenue.**

Analysis and modeling of previous bookings provide **insights into the drivers** of cancellations and **identify the characteristics** of bookings most **likely to be canceled.**

These insights that can inform strategies to **minimize unfilled rooms. Maximizing revenue,** and **improving operational efficiency.**

·02· Approach & Methodology

Data Available

Feature	Description
Country	Country of origin. Categories are represented in the ISO 3155–3:2013 format.
Deposit Type	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories No-Deposit, Non-Refund and Refundable
Lead Time	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
Total of Special Request	Number of special requests made by the customer (e.g. twin bed or high floor).
Average Daily Rate(ADR)	Calculated by dividing the sum of all lodging transactions by the total number of staying nights.
Market Segment	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
Arrival Date Day of Month	Day of the month of the arrival date.
Arrival Date Week Number	Week number of the arrival date.
Stays In Week Nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel. Calculated by counting the number of week nights from the total number of nights.

119,390

Records

32

Features

2015-2017

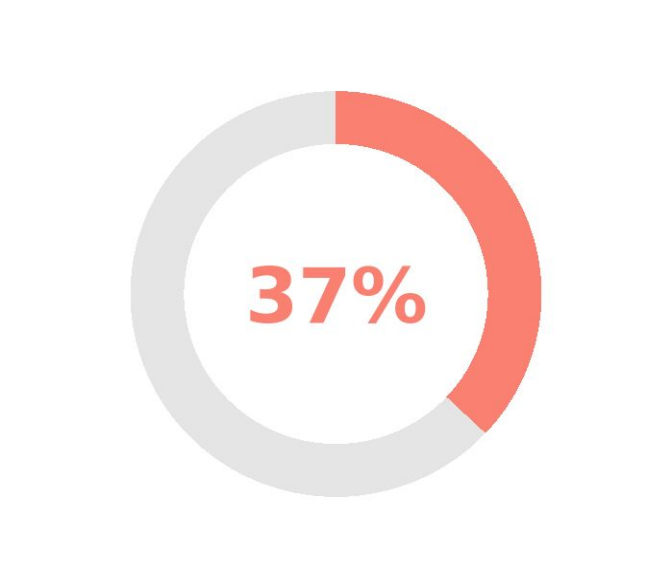
Time Range

Approach

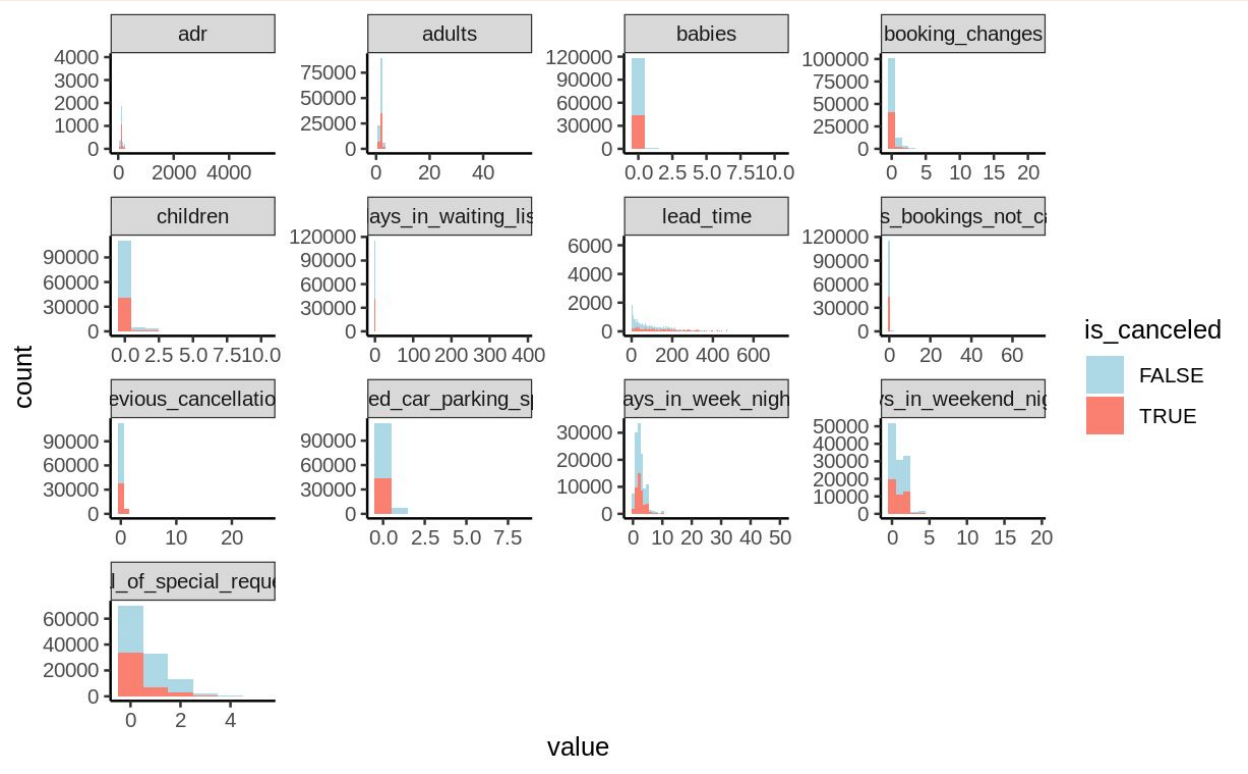
1. EDA
2. Data Cleaning
3. Logistic Regression
4. Random Forest

·03· Data Cleaning

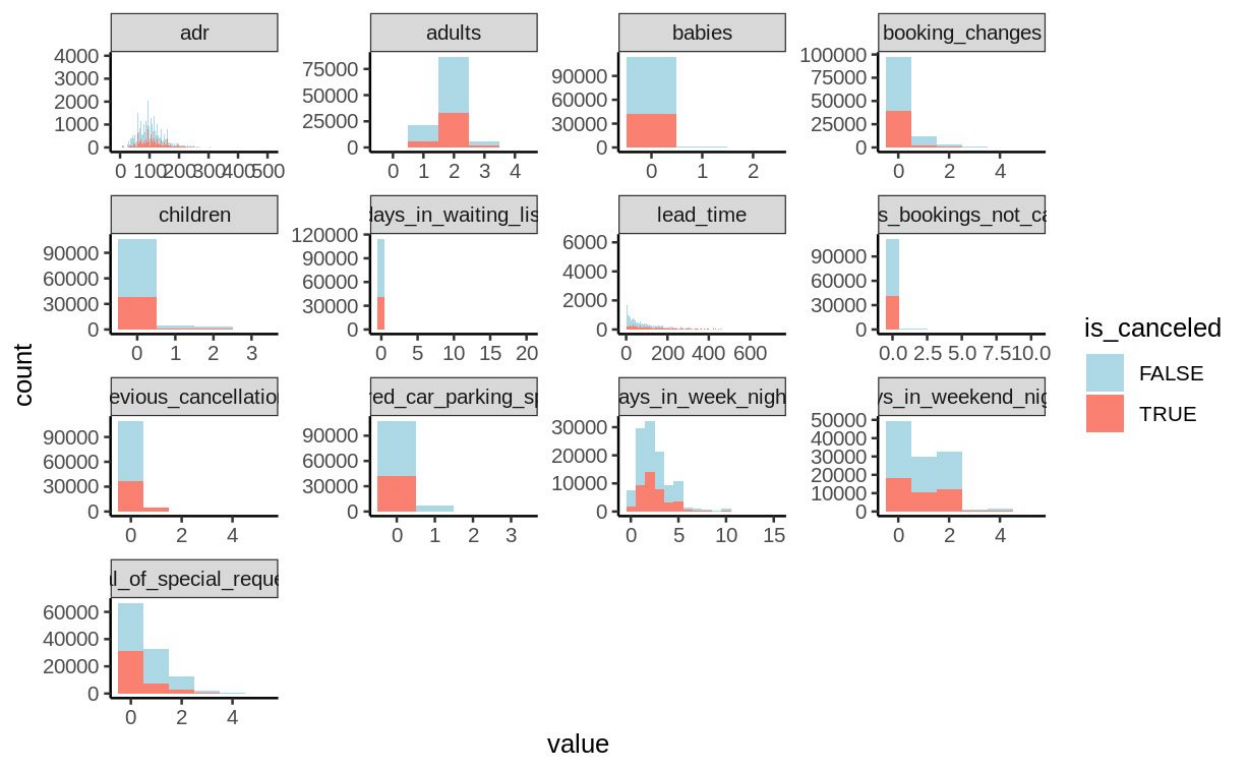
Overall Cancellation Rate



Distribution of Quantitative Variables



Distribution of Quantitative Variables



Distribution of Categorical Variables

Predictor	Predictor Value	Number of Rows	Relative Proportion		Predictor	Predictor Value	Number of Rows	Relative Proportion
hotel	City Hotel	75755	66%		deposit_type	No Deposit	102304	89%
hotel	Resort Hotel	39367	34%		deposit_type	Non Refund	12667	11%
arrival_date_month	August	13771	12%		deposit_type	Refundable	151	0%
arrival_date_month	July	12536	11%		distribution_channel	TA/TO	94180	81.81%
arrival_date_month	May	11222	10%		distribution_channel	Direct	14435	12.54%
arrival_date_month	October	10606	9.21%		distribution_channel	Corporate	6314	5.48%
arrival_date_month	April	10587	9.20%		distribution_channel	GDS	192	0.17%
arrival_date_month	June	10482	9.11%		distribution_channel	Undefined	1	0.00%

Removing Features

- Time series predictors removed
- Reservation status was related directly to cancellation status
- Out of the remaining columns, none had a generalized VIF >= 5.

Predictor	GVIF	Df	GVIF^(1/(2*Df))
is_repeated_guest	1.42E+00	1	1.193426
lead_time	1.46E+00	1	1.208154
stays_in_weekend_nights	1.29E+00	1	1.137094
stays_in_week_nights	1.44E+00	1	1.198095
adults	1.31E+00	1	1.146497
children	2.08E+00	1	1.441182
babies	1.03E+00	1	1.013874
previous_cancellations	1.42E+00	1	1.191814
previous_bookings_not_canceled	1.57E+00	1	1.25347
booking_changes	1.06E+00	1	1.027931
days_in_waiting_list	1.01E+00	1	1.006472

After Cleaning

115,103	~4000	96.4%	25
Data Points Remaining	Records Removed	Data Retained	Features

·04· Results

Data Preparation

- Before fitting the logistic regression, we performed an 80/20 train-test split, using the test set to prevent overfitting.
- We included all numerical variables in the data following previous data cleaning steps. These variables include "lead time", "total special requests", "number of adults", and others.
- We also included all categorical variables after data cleaning, such as "meal", "customer type", and "reserved room type", among others.

Simple Logistic Regression

- We fitted the model on the training set and demonstrated the out-of-sample prediction results on the test set.
- The simple logistic regression already did a good job in predicting the booking cancellations as well as identifying those users who did not cancel their bookings.

Sensitivity	83.62%
Specificity	95.01%
FP Rate	4.99%
Precision	90.52%
Accuracy	90.87%

Logistic Regression with Ridge

- We further explored the potential of using regularization to improve the model's performance.
- We tried Ridge first and used cross validation with root mean squared error as the loss function to select the best penalizing parameter, lambda.
- Ridge improved some of the metrics but not all.

	Simple Logit	Logit with Ridge
Sensitivity	83.62%	78.7%
Specificity	95.01%	95.7%
FP Rate	4.99%	4.3%
Precision	90.52%	91.2%
Accuracy	90.87%	89.5%

Logistic Regression with Lasso

- We also implemented Lasso regression, using cross-validation to determine the optimal lambda parameter.
- The results show that Ridge and Lasso improved the model performance in different sets of metrics.
- Lasso appears to be more suitable for our application as it exhibits superior performance in predicting booking cancellations.

	Simple Logit	Logit with Ridge	Logit with Lasso
Sensitivity	83.62%	78.7%	89.29%
Specificity	95.01%	95.7%	93.52%
FP Rate	4.99%	4.3%	6.48%
Precision	90.52%	91.2%	88.71%
Accuracy	90.87%	89.5%	91.99%

Logistic Regression with Elastic Net

- We decided to combine Ridge and Lasso techniques to see if we could achieve the highest performing model.
- We partitioned the parameter space [0,1] into ten segments, increasing by increments of 0.1 to adjust the weighting parameter, alpha.
- Cross-validation was then used to determine the optimal pair of lambda and alpha.

	Simple Logit	Logit with Ridge	Logit with Lasso	Logit with Elastic Net
Sensitivity	83.62%	78.7%	89.29%	89.29%
Specificity	95.01%	95.7%	93.52%	93.51%
FP Rate	4.99%	4.3%	6.48%	6.49%
Precision	90.52%	91.2%	88.71%	88.69%
Accuracy	90.87%	89.5%	91.99%	91.98%

Random Forest

- A random forest model was built to predict hotel cancellations.
- Model was trained on 70% of the cleaned data. The remaining 30% of the data was used for testing.
- The 2 hyperparameters tuned for the model were mtry and ntree.

Random Forest - mtry

- Tested mtry values of 1 through 10.
- These models were all built using ntree = 300.
- mtry = 8 had the best test error, so this was selected for the final model.

Random Forest - mtry

mtry	Train Error	Test Error
1	25.11	24.72
2	17.86	17.29
3	15.96	15.74
4	15.08	14.8
5	14.42	14.17
6	14.09	13.85
7	13.95	13.81
8	13.9	13.62
9	13.9	13.65
10	13.97	13.8

Random Forest - ntree

- Tested ntree values of 500, 1000, and 1500.
- These models were all built using mtry = 8.
- ntree = 300 had the best test error, so this was selected for the final model.

Random Forest - ntree

mtry	ntree	Train Error	Test Error
8	300	13.9	13.62
8	500	13.89	13.7
8	1000	13.85	13.66
8	1500	13.85	13.69

Random Forest - Final Model

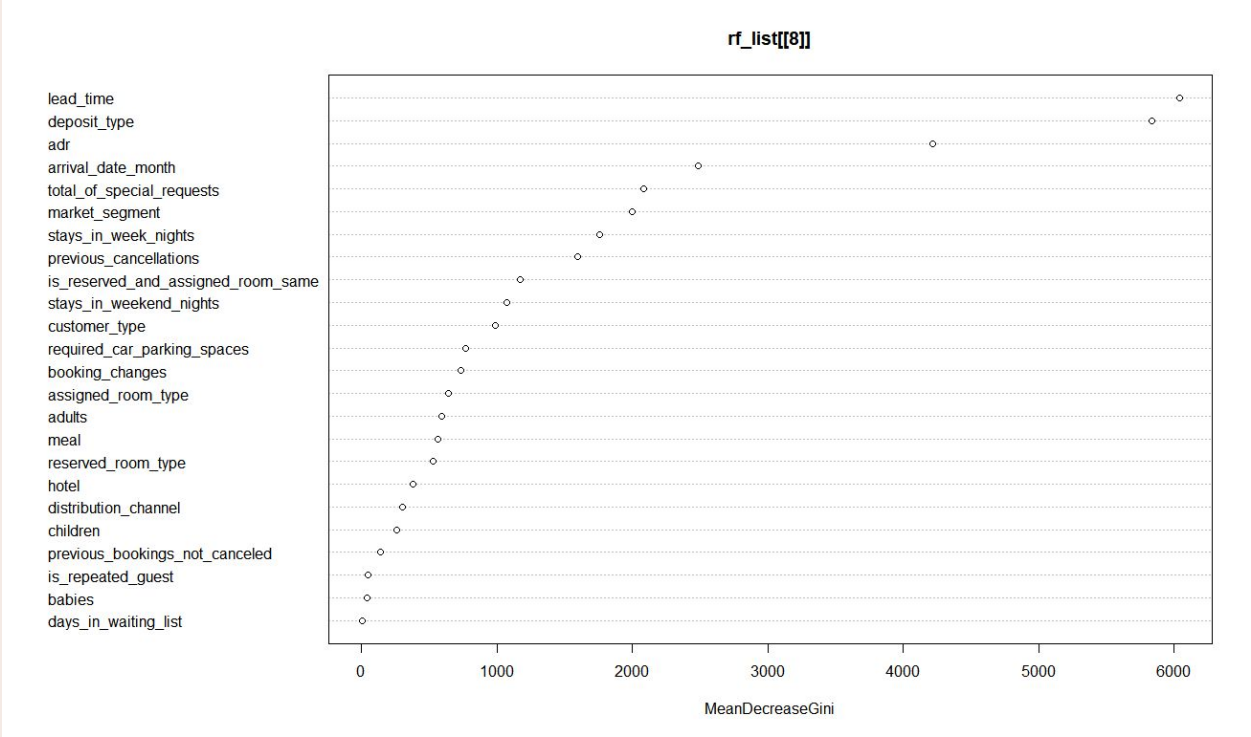
Training data confusion matrix

		Actual Values		Class Error
		0	1	
Predicted Values	0	47048	4294	0.0836
	1	6907	22324	0.2363

Testing data confusion matrix

		Actual Values		Class Error
		0	1	
Predicted Values	0	20166	1837	0.0835
	1	2867	9660	0.2289

Random Forest - Final Model



Random Forest - Final Model

- The final model had the following performance:
  - Sensitivity = 77.11%
  - Specificity = 91.65%
  - Accuracy = 86.38%
- Based on the model's variable importance score, these were the most important features of the data:
  - Lead time
  - Deposit type
  - Average daily rate

·05· MODEL COMPARISONS

Logistic Regression vs. Random Forest

	Logistic Regression	Random Forest
Sensitivity	83.62%	77.11%
Specificity	95.01%	91.65%
Accuracy	90.87%	86.38%

·06· APPLICATIONS

Logistic Regression

- Resource Allocation: Hotels can prioritize bookings with high cancellation probabilities, ensuring efficient resource distribution.
- Preventative Action: Early warning system capabilities allow hotels to take proactive measures to minimize cancellations and optimize room occupancy
- Segmentation: Logistic regression allows the segmentation of clients based on their propensity to cancel reservations.
- Personalization: Hotels can tailor services and marketing strategies for different customer groups, enhancing customer satisfaction and loyalty.
- Incentives: Providing targeted incentives to clients who are likely to cancel can improve retention rates and reduce cancellations.

Random Forest

- Marketing Focus: Hotels can concentrate their marketing efforts on elements that lead to confirmed reservations, improving conversion rates.
- Optimizing Overbooking: Using data from random forests, hotels can calculate the ideal level of overbookings to minimize the impact of cancellations on room occupancy.
- Combining Demand Information: Random forest forecasts, combined with demand data, enable dynamic pricing strategies.
- Real-Time Pricing: Hotels can modify prices in real-time based on the likelihood of cancellations, maximizing revenue and occupancy.
- Data-Driven Decisions: Utilizing insights from random forests empowers hotels to make informed business decisions.

Ensemble Model

- Operational Management: Hotels can make data-driven decisions regarding resource allocation, staffing, and marketing strategies.
- Minimizing Disruptions: Understanding cancellation risks allows hotels to prepare for potential disruptions and maintain smooth operations.
- Tailored Customer Engagement: Ensembles enable hotels to identify customers more likely to cancel and engage them with personalized incentives.
- Enhanced Customer Service: Proactive outreach to potential cancellations can lead to improved customer satisfaction and retention.
- Optimizing Revenue: Effective client retention strategies lead to increased bookings and revenue for the hotel.

·07· FUTURE WORKS AND CONCLUSION

Future works and Conclusion

Future Works

- Explore time series analysis for cancellation trends.
- Improve predictions with ensemble models and SMOTE.
- Utilize sentiment analysis of customer reviews for insights.
- Investigate model interpretation for robust conclusions.
- Pair the model with pricing strategy analysis for optimal overbooking decisions.

Future works and Conclusion

Conclusion

- Despite real-world data challenges, our predictive classification models offer significant value to the hospitality industry.
- With ~90% accuracy, our models highlight lead time, deposit type, and average daily rate as crucial factors for predicting hotel cancellations

References

- Andriawan, Z. A., Purnama, S. R., Darmawan, A. S., Wibowo, A., Sugiharto, A., Wijayanto, F. et al. (2020), Prediction of hotel booking cancellation using crisp-dm, in ‘2020 4th International Conference on Informatics and Computational Sciences (ICICoS)’, IEEE, pp. 1–6.

- Antonio, N., de Almeida, A. & Nunes, L. (2017), ‘Predicting hotel booking cancellations to decrease uncertainty and increase revenue’, Tourism & Management Studies 13(2), 25– 39.

THANK YOU!