Predicting Hotel Cancelations

Capstone Project

Kerem Onipede

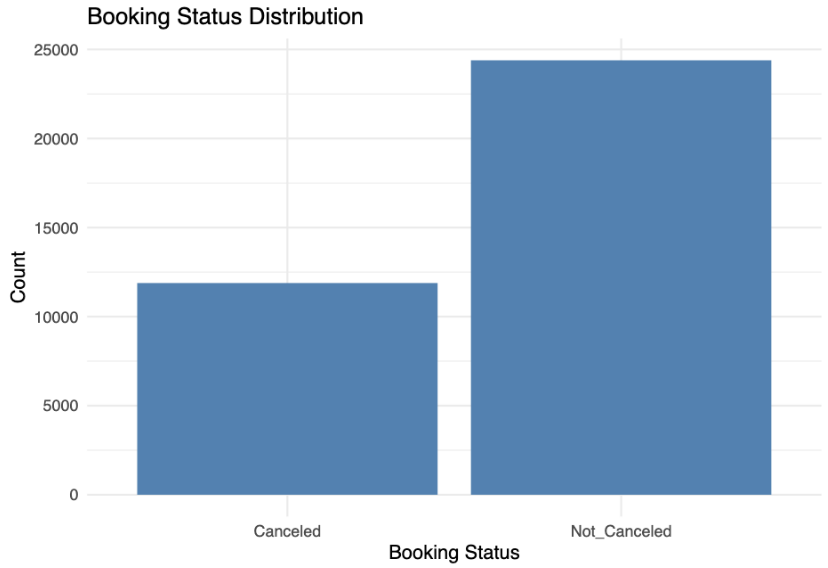
Data Set

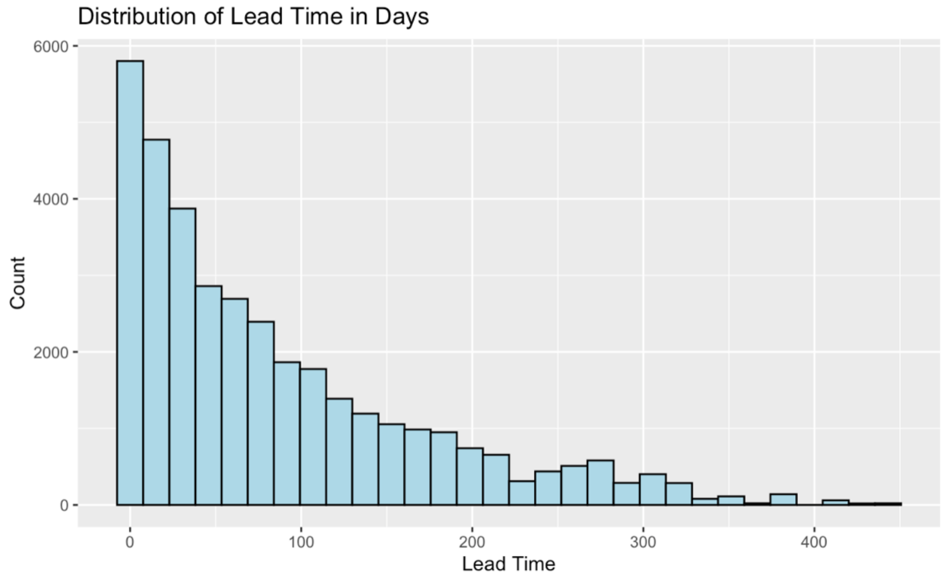
- From 1 hotel in Lisbon, Portugal, 2017-2018
- 36,275 observations
- 16 variables
    - **no_of_adults** – integer
    - **no_of_children** – integer
    - **no_of_weekend_nights** – integer
    - **no_of_week_nights** – integer
    - **type_of_meal_plan** – factor
    - **required_car_parking_space** – integer
    - **room_type_reserved** – factor
    - **arrival_month** – factor
    - **market_segment_type** – factor
    - **repeated_guest** – integer
    - **no_of_previous_cancellations** – integer
    - **no_of_previous_bookings_not_canceled** – integer
    - **no_of_special_requests** – integer
    - **booking_status** – factor
    - **lead_time** – numeric
    - **avg_price_per_room** – numeric

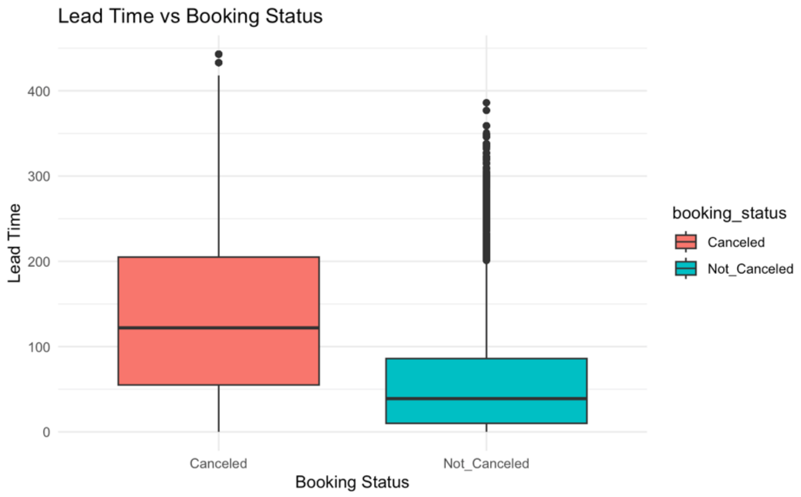
Objective

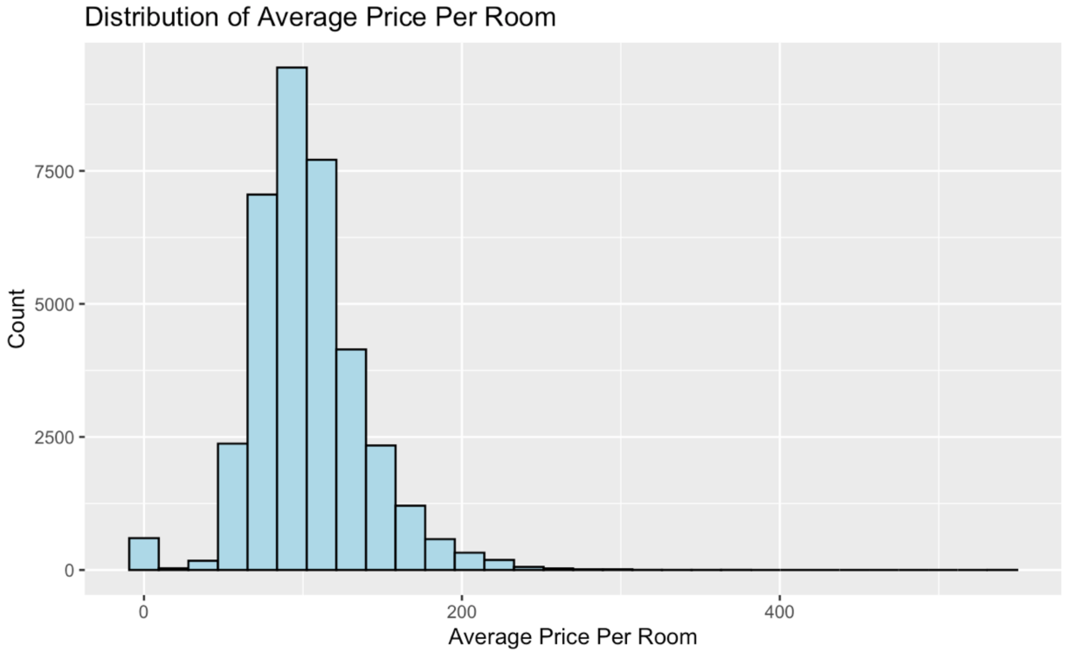
- Can we predict how likely a booking is to be canceled?
- What factors affect how likely a booking is to be canceled?

Exploratory Analysis

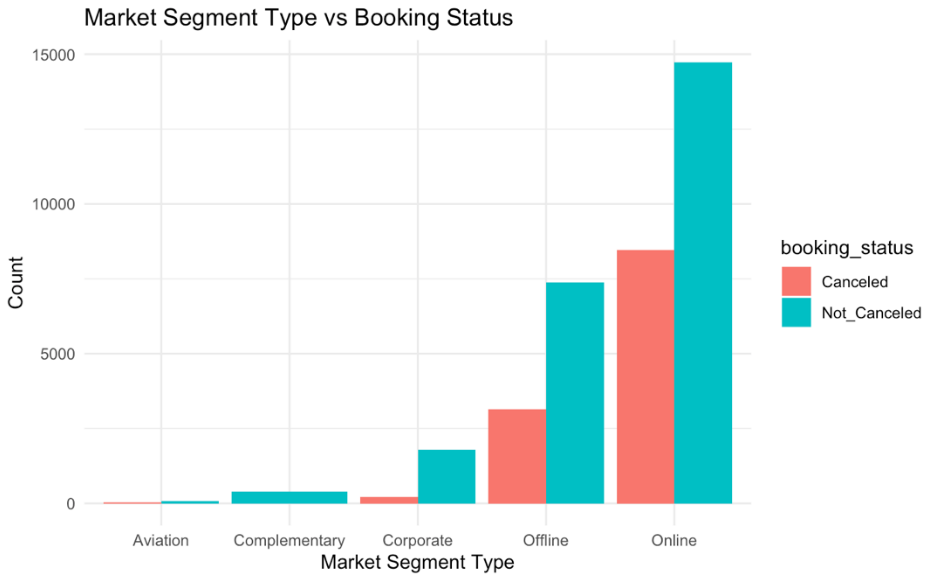


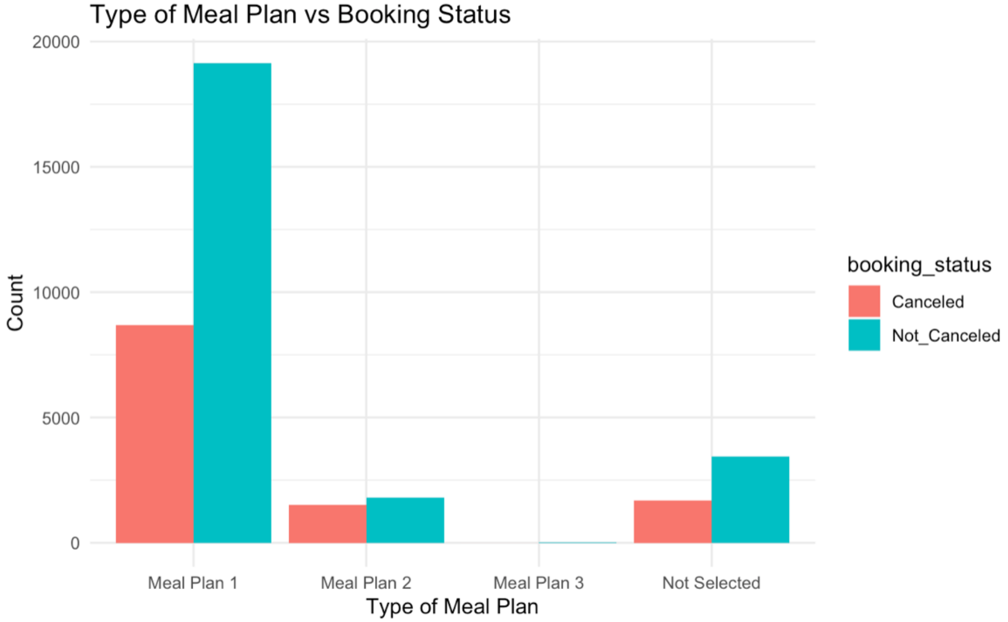


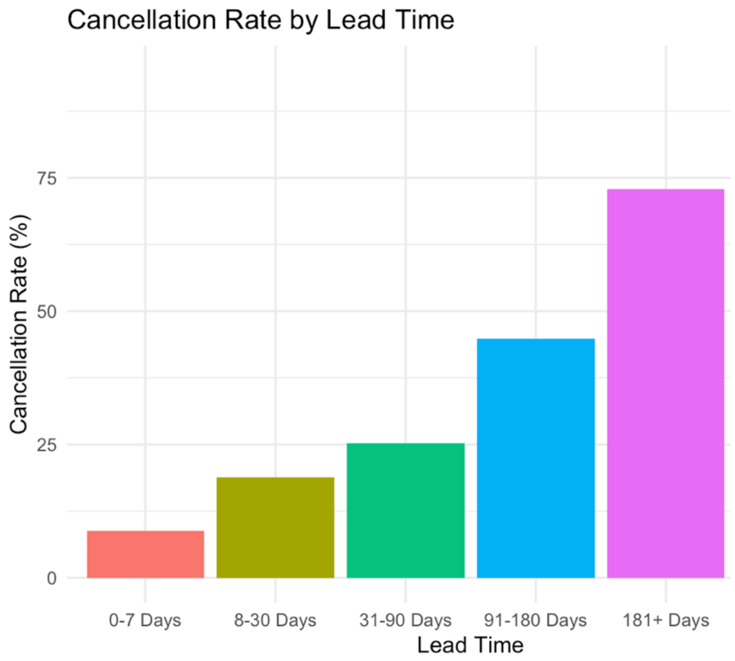


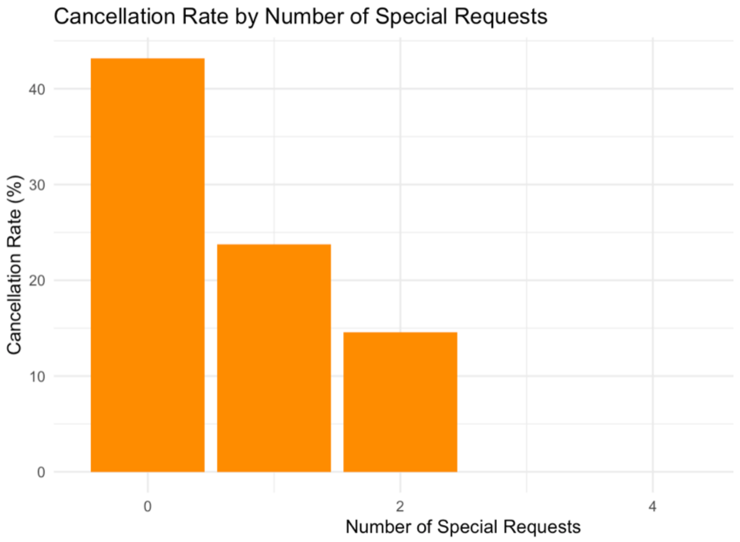


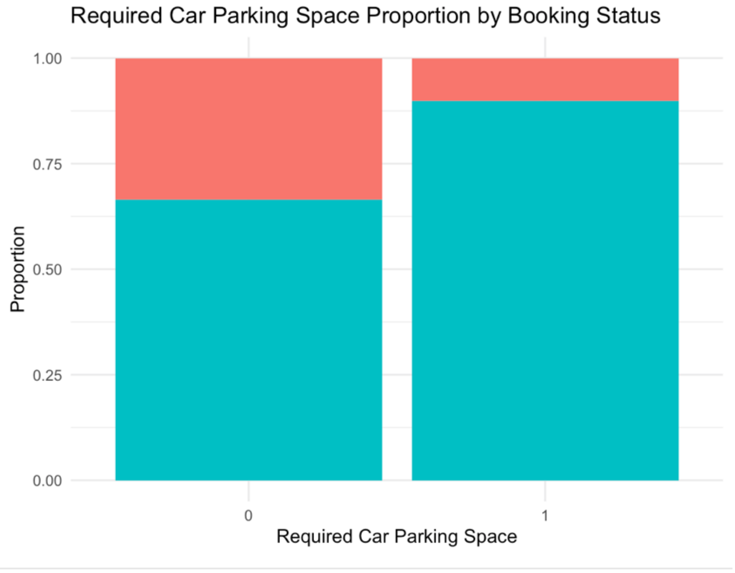
Exploratory Analysis






Cancellation Rate by Lead Time
Cancellation Rate (%)
75
50
25
0
0-7 Days
8-30 Days
31-90 Days
91-180 Days
181+ Days
Lead Time


Cancellation Rate by Number of Special Requests
Cancellation Rate (%)
40
30
20
10
0
0
2
4
Number of Special Requests

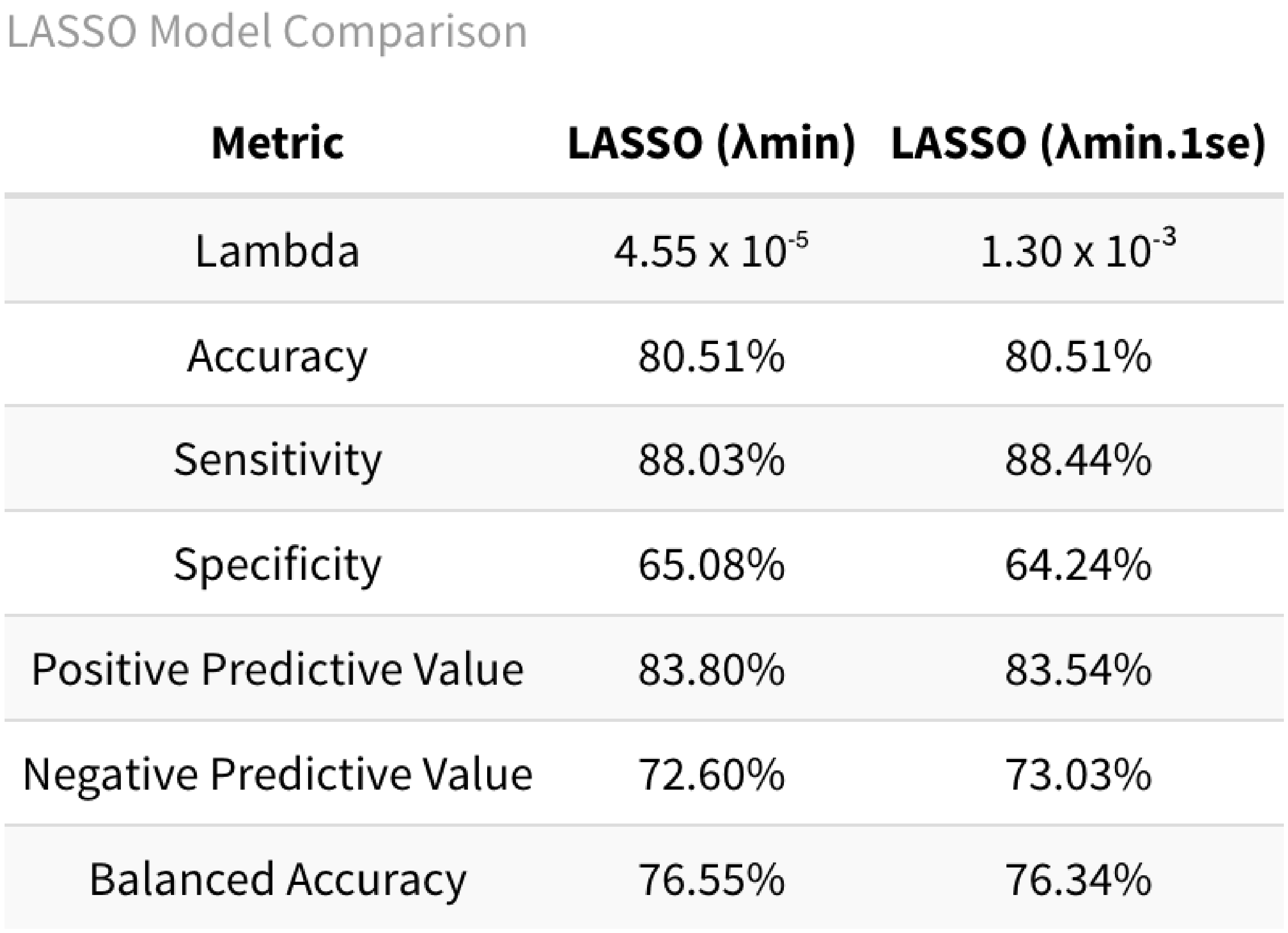

Required Car Parking Space Proportion by Booking Status
Proportion
1.00
0.75
0.50
0.25
0.00
0
1
Required Car Parking Space

Preparing the data

- **Target Variable Transformation**
  - Convert booking_status to a binary factor (1 = Canceled, 0 = Not Canceled)
- **Feature Engineering & Cleaning**
  - Drop non-predictive columns (e.g., Booking_ID, arrival_date) and convert date to month
  - Convert arrival_month and other categorical predictors (type_of_meal_plan, room_type_reserved, market_segment_type) to factors
- **Handling Skewed Variables**
  - Apply log transformation to lead_time and avg_price_per_room
- **Data Scaling**
  - Center and scale the predictors to normalize the data for improved model performance
- **Data Partitioning**
  - Split data into training (80%) and testing (20%) sets for model evaluation

LASSO

- Chose lambda using 10 fold CV
- Tried lambda.min and lambda.1se
- Similar performance
- Chose lambda.1se for simpler model

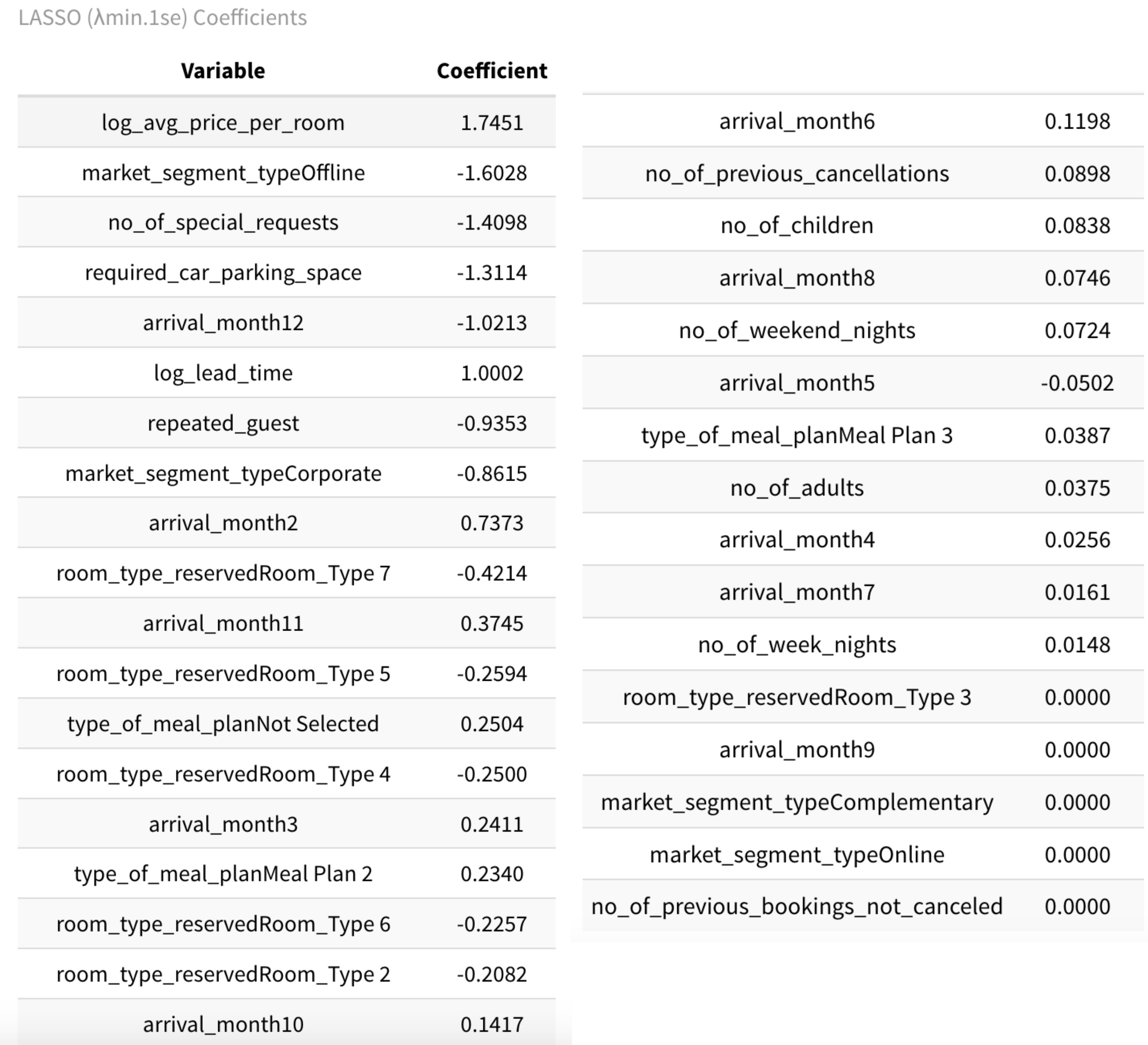


LASSO Model Comparison

Metric	LASSO (λmin)	LASSO (λmin.1se)
Lambda	4.55 x 10^{-5}	1.30 x 10^{-3}
Accuracy	80.51%	80.51%
Sensitivity	88.03%	88.44%
Specificity	65.08%	64.24%
Positive Predictive Value	83.80%	83.54%
Negative Predictive Value	72.60%	73.03%
Balanced Accuracy	76.55%	76.34%

LASSO Coefficients

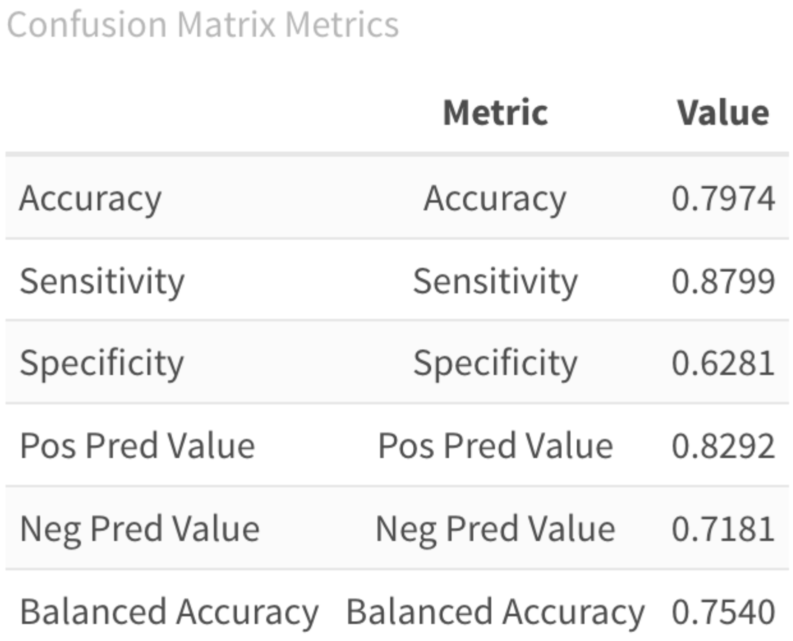
- The LASSO approach applies a penalty that shrinks less relevant predictors to zero, leaving only those with the most predictive power.
- In this table, coefficients are sorted by absolute value. Higher magnitude indicates a stronger impact on whether a booking is canceled.
- Positive coefficients suggest an increased likelihood of cancellation, while negative coefficients indicate a decreased likelihood



LASSO (λmin.1se) Coefficients

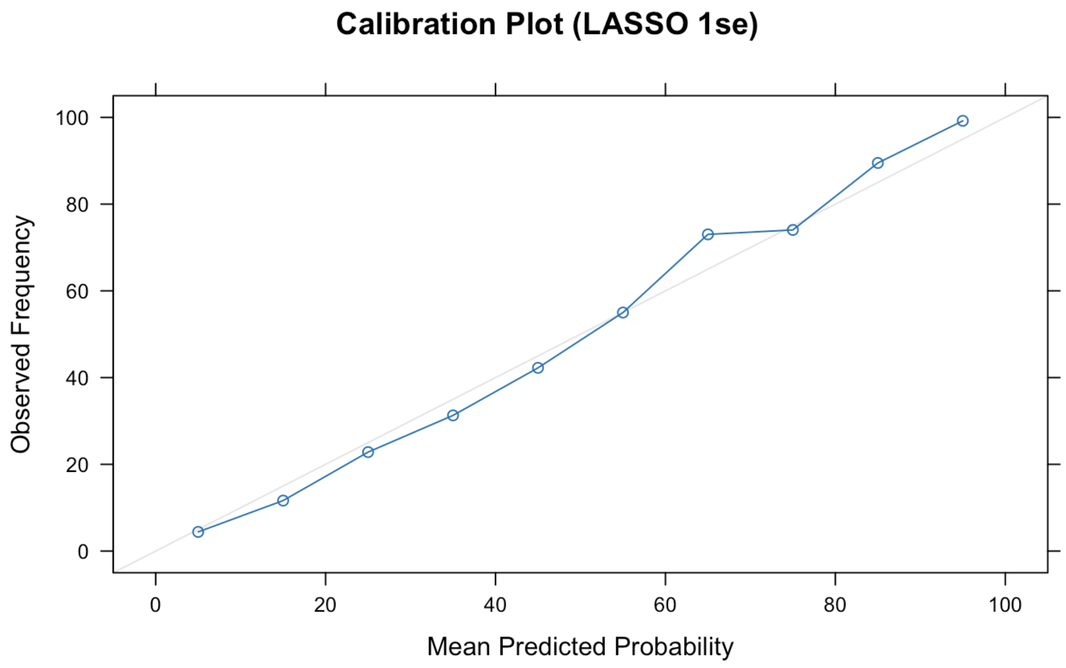
Variable	Coefficient
log_avg_price_per_room	1.7451
market_segment_typeOffline	-1.6028
no_of_special_requests	-1.4098
required_car_parking_space	-1.3114
arrival_month12	-1.0213
log_lead_time	1.0002
repeated_guest	-0.9353
market_segment_typeCorporate	-0.8615
arrival_month2	0.7373
room_type_reservedRoom_Type 7	-0.4214
arrival_month11	0.3745
room_type_reservedRoom_Type 5	-0.2594
type_of_meal_planNot Selected	0.2504
room_type_reservedRoom_Type 4	-0.2500
arrival_month3	0.2411
type_of_meal_planMeal Plan 2	0.2340
room_type_reservedRoom_Type 6	-0.2257
room_type_reservedRoom_Type 2	-0.2082
arrival_month10	0.1417
arrival_month6	0.1198
no_of_previous_cancellations	0.0898
no_of_children	0.0838
arrival_month8	0.0746
no_of_weekend_nights	0.0724
arrival_month5	-0.0502
type_of_meal_planMeal Plan 3	0.0387
no_of_adults	0.0375
arrival_month4	0.0256
arrival_month7	0.0161
no_of_week_nights	0.0148
room_type_reservedRoom_Type 3	0.0000
arrival_month9	0.0000
market_segment_typeComplementary	0.0000
market_segment_typeOnline	0.0000
no_of_previous_bookings_not_canceled	0.0000

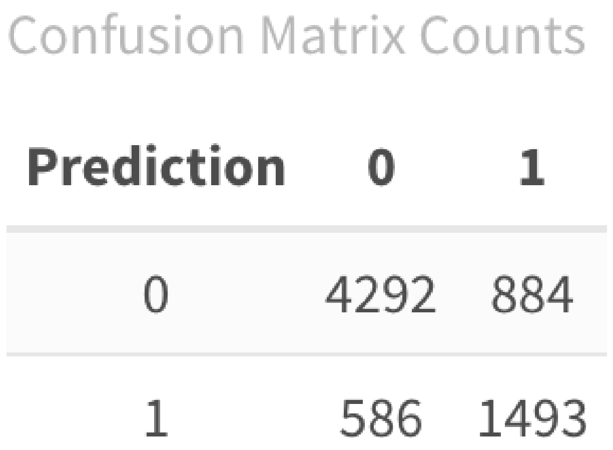
LASSO Model



Confusion Matrix Metrics

	Metric	Value
Accuracy	Accuracy	0.7974
Sensitivity	Sensitivity	0.8799
Specificity	Specificity	0.6281
Pos Pred Value	Pos Pred Value	0.8292
Neg Pred Value	Neg Pred Value	0.7181
Balanced Accuracy	Balanced Accuracy	0.7540



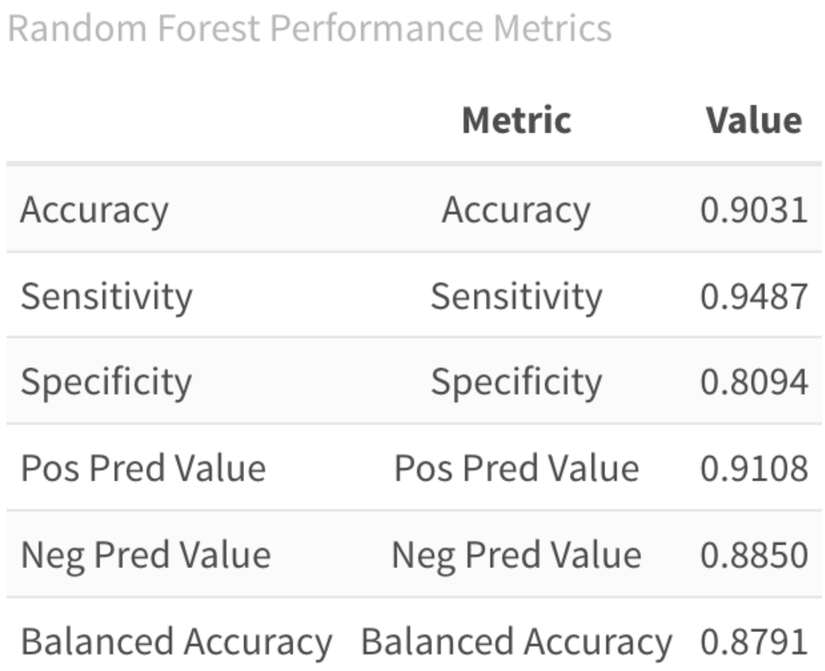


Confusion Matrix Counts

Prediction	0	1
0	4292	884
1	586	1493

Random Forest

- Tuned mtry using F1 score
- Best mtry = 4



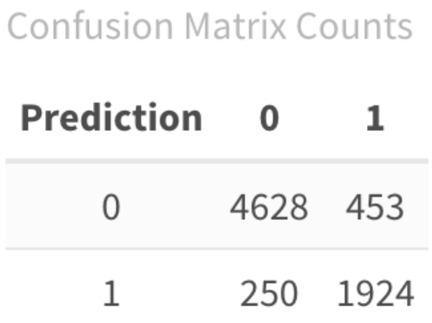
Random Forest Performance Metrics

	Metric	Value
Accuracy	Accuracy	0.9031
Sensitivity	Sensitivity	0.9487
Specificity	Specificity	0.8094
Pos Pred Value	Pos Pred Value	0.9108
Neg Pred Value	Neg Pred Value	0.8850
Balanced Accuracy	Balanced Accuracy	0.8791

mtry	F1-score	Accuracy
2	0.8000	0.8815
3	0.8343	0.8973
4	**0.8457**	**0.9031**
5	0.8451	0.9023
6	0.8440	0.9016
7	0.8453	0.9023
8	0.8457	0.9024
9	0.8427	0.9005

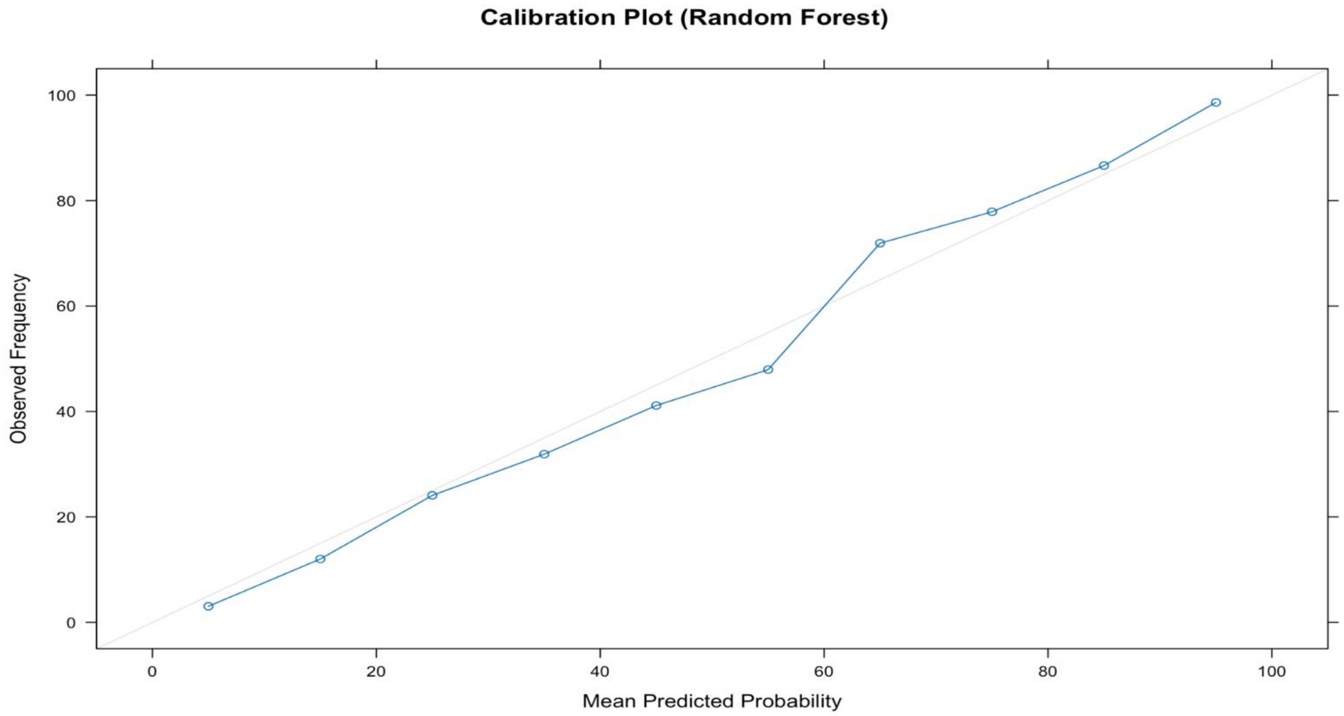
Random Forest

- **Random Forest Results**
- **Calibration Plot**
- Overpredict from 55-65
- Under predict from 65-80



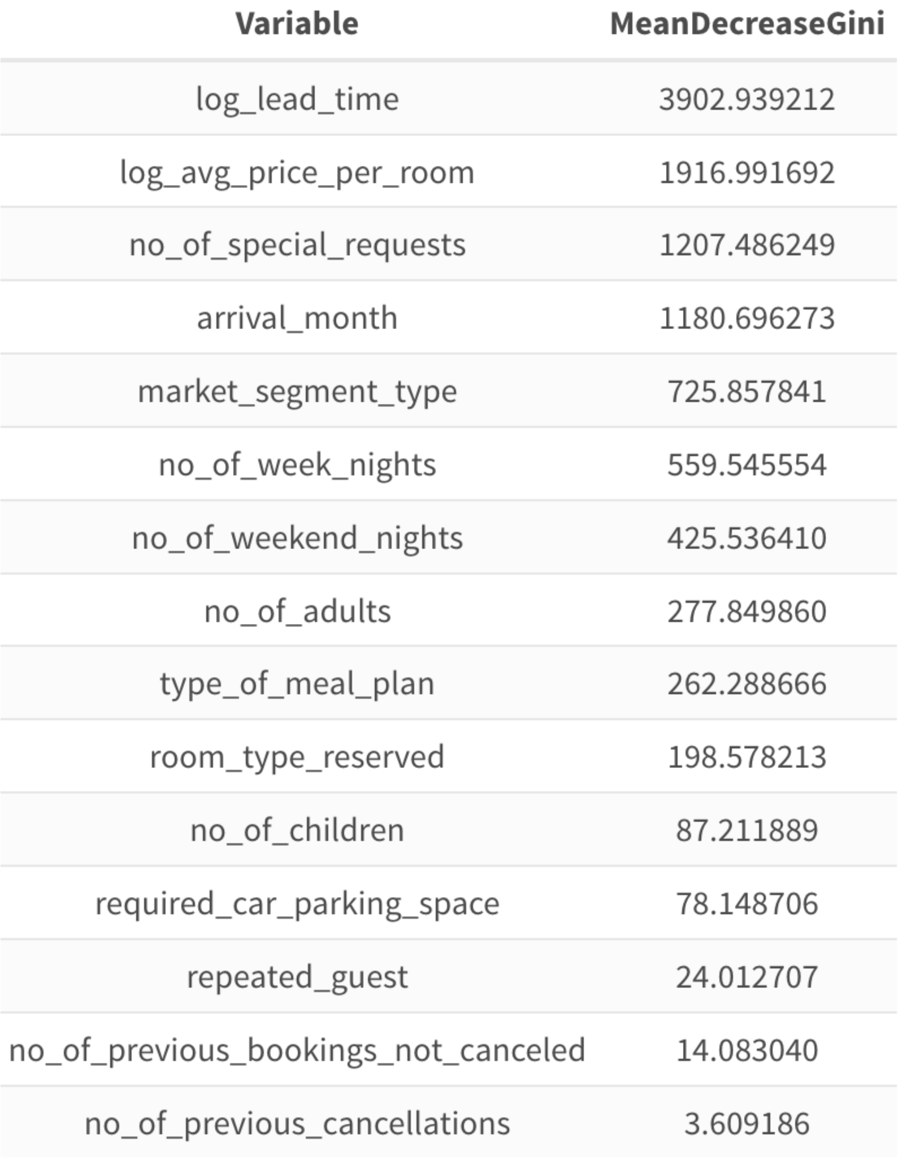
Confusion Matrix Counts

Prediction	0	1
0	4628	453
1	250	1924



Calibration Plot (Random Forest)

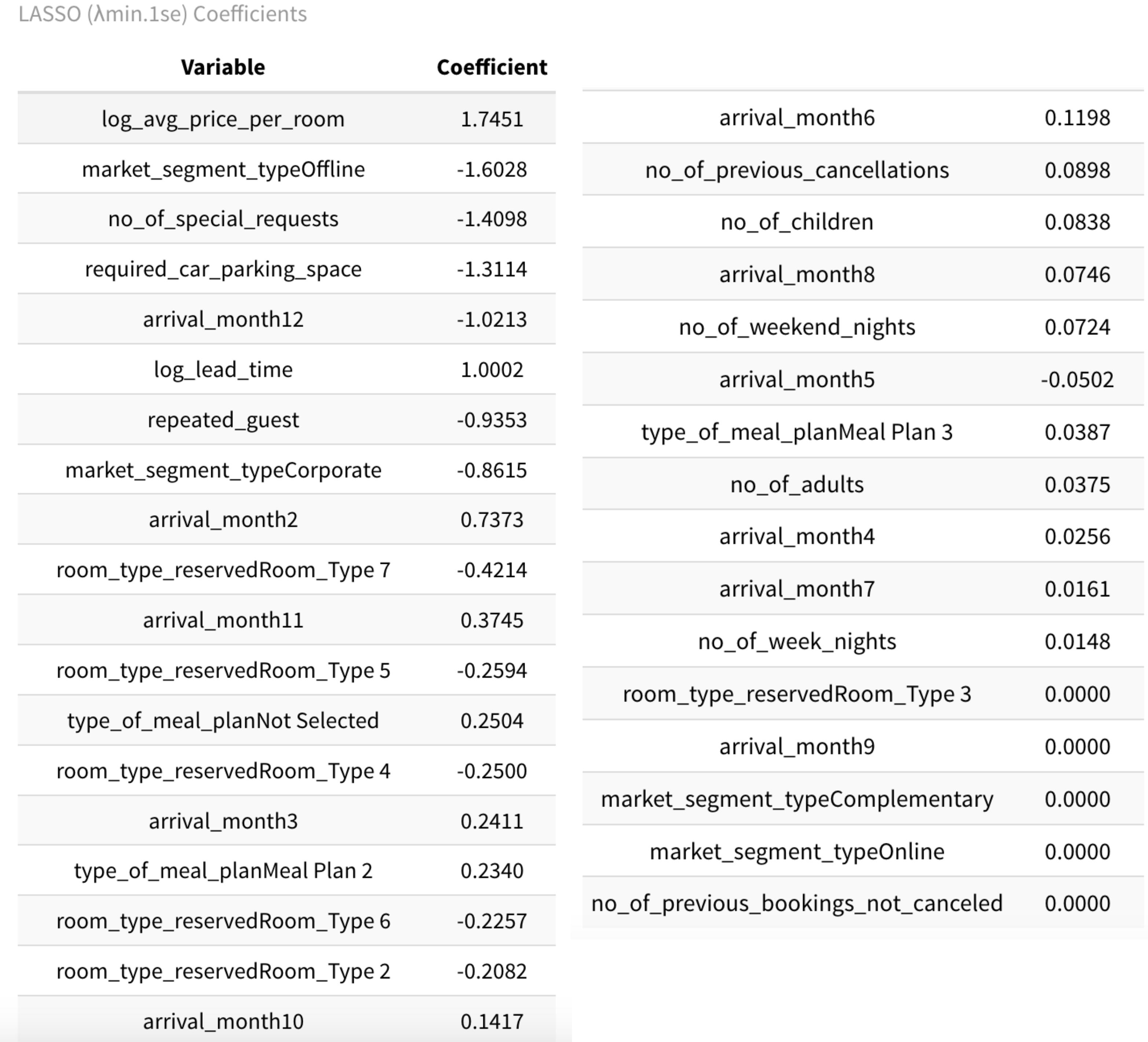
RF Features



Variable	MeanDecreaseGini
log_lead_time	3902.939212
log_avg_price_per_room	1916.991692
no_of_special_requests	1207.486249
arrival_month	1180.696273
market_segment_type	725.857841
no_of_week_nights	559.545554
no_of_weekend_nights	425.536410
no_of_adults	277.849860
type_of_meal_plan	262.288666
room_type_reserved	198.578213
no_of_children	87.211889
required_car_parking_space	78.148706
repeated_guest	24.012707
no_of_previous_bookings_not_canceled	14.083040
no_of_previous_cancellations	3.609186

LASSO Coefficients

- The LASSO approach applies a penalty that shrinks less relevant predictors to zero, leaving only those with the most predictive power.
- In this table, coefficients are sorted by absolute value—higher magnitude indicates a stronger impact on whether a booking is canceled.
- Positive coefficients suggest an increased likelihood of cancellation, while negative coefficients indicate a decreased likelihood



LASSO (λmin.1se) Coefficients

Variable	Coefficient
log_avg_price_per_room	1.7451
market_segment_typeOffline	-1.6028
no_of_special_requests	-1.4098
required_car_parking_space	-1.3114
arrival_month12	-1.0213
log_lead_time	1.0002
repeated_guest	-0.9353
market_segment_typeCorporate	-0.8615
arrival_month2	0.7373
room_type_reservedRoom_Type 7	-0.4214
arrival_month11	0.3745
room_type_reservedRoom_Type 5	-0.2594
type_of_meal_planNot Selected	0.2504
room_type_reservedRoom_Type 4	-0.2500
arrival_month3	0.2411
type_of_meal_planMeal Plan 2	0.2340
room_type_reservedRoom_Type 6	-0.2257
room_type_reservedRoom_Type 2	-0.2082
arrival_month10	0.1417
arrival_month6	0.1198
no_of_previous_cancellations	0.0898
no_of_children	0.0838
arrival_month8	0.0746
no_of_weekend_nights	0.0724
arrival_month5	-0.0502
type_of_meal_planMeal Plan 3	0.0387
no_of_adults	0.0375
arrival_month4	0.0256
arrival_month7	0.0161
no_of_week_nights	0.0148
room_type_reservedRoom_Type 3	0.0000
arrival_month9	0.0000
market_segment_typeComplementary	0.0000
market_segment_typeOnline	0.0000
no_of_previous_bookings_not_canceled	0.0000

Feature Importance Comparison

- **Interpretation Approach**
  - **Random Forest** uses **Mean Decrease in Gini** to measure how each feature reduces impurity across all splits. Larger values mean the feature is more influential in splitting the data correctly.
  - **LASSO** relies on **coefficient magnitude** to gauge importance. Larger absolute coefficients indicate stronger influence on the outcome, and the sign (+/−) shows whether the feature increases or decreases the likelihood of cancellation.
- **Top Features**
  - **Random Forest** highlights:
    - **Log Lead Time** as the most critical predictor, followed by **Log Avg. Price per Room** and **No. of Special Requests**.
    - Arrival month and market segment also rank high but appear slightly less influential than lead time or price.
  - **LASSO** emphasizes:
    - **Market Segment Type (e.g., Offline)** with a large (negative) coefficient, meaning offline bookings reduce cancellation likelihood.
    - **Log Avg. Price per Room** and **No. of Special Requests** also show high coefficients, suggesting strong effects on cancellation probability.
- **Overlap & Differences**
  - Both methods underscore **lead time**, **average price**, and **special requests** as critical.
  - **Random Forest** ranks features purely by predictive splitting power, whereas **LASSO** provides directionality—positive or negative impact on cancellation.
  - Certain categorical variables (e.g., **market segment**, **meal plan type**) may show up strongly in LASSO due to their coefficients but are spread out in Random Forest importance.
- **Key Takeaway**
  - **Log Lead Time** and **Avg. Price** are universally important predictors.
  - **LASSO** more explicitly reveals **direction** (increase vs. decrease cancellation likelihood).
  - **Random Forest** highlights **overall predictive power** and interactions, making it robust to nonlinear relationships.

Feature Importance Analysis

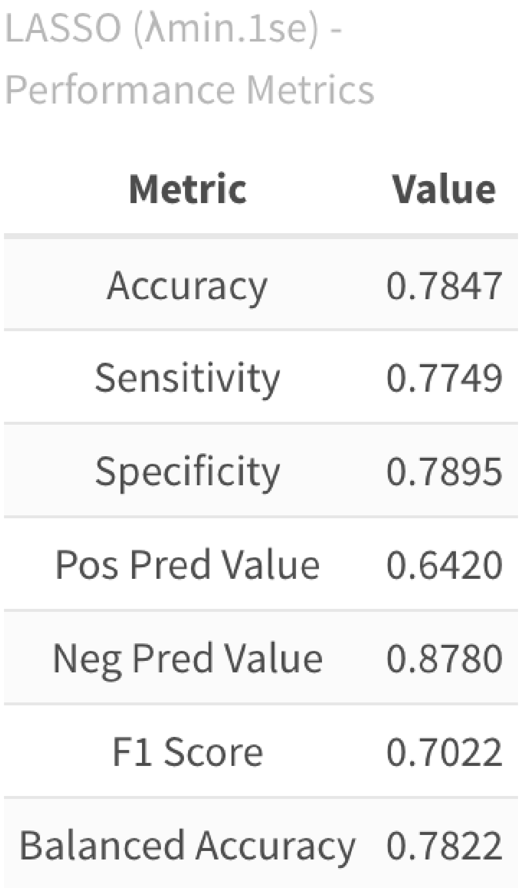
- **log_lead_time**
- *What it is:* Log-transformed days between booking and arrival.
- *Interpretation:* A positive coefficient indicates that longer lead times increase the likelihood of cancellation.
- **no_of_special_requests**
- *What it is:* Count of additional requests made by the guest (e.g., extra pillows, room preferences).
- *Interpretation:* A negative coefficient suggests that more special requests are linked to a lower probability of cancellation.
- **log_avg_price_per_room**
- *What it is:* Log-transformed average price per room.
- *Interpretation:* A positive effect shows that higher room prices are associated with a higher cancellation risk.
- **market_segment_type**
- *What it is:* Categorical indicator of the booking channel (e.g., online, offline).
- *Interpretation:* Certain segments (for instance, offline bookings) tend to lower cancellation likelihood.
- **arrival_month**
- *What it is:* The month in which the guest is scheduled to arrive.
- *Interpretation:* Variations here reveal seasonal trends; some months have higher cancellation probabilities than others.

Feature Importance

- **repeated_guest**
- *What it is:* Binary indicator showing if the guest is a repeat customer.
- *Interpretation:* A negative coefficient implies that repeat guests are less likely to cancel.
- **no_of_previous_cancellations**
- *What it is:* Number of times the guest has canceled in the past.
- *Interpretation:* A positive coefficient indicates that a history of cancellations increases the risk of future cancellations.
- **no_of_previous_bookings_not_canceled**
- *What it is:* Count of the guest’s previous bookings that were successfully completed.
- *Interpretation:* A negative effect suggests that a record of completed bookings reduces the probability of cancellation.
- **no_of_weekend_nights**
- *What it is:* The number of weekend nights included in the booking.
- *Interpretation:* The coefficient shows that the number of weekends nights slightly increases the risk of cancellation.
- **type_of_meal_plan**
- *What it is:* Categorical variable indicating the chosen meal plan (e.g., breakfast only, full board).
- *Interpretation:* Differences in meal plan selections may reflect varying levels of commitment or guest expectations, influencing cancellation risk.

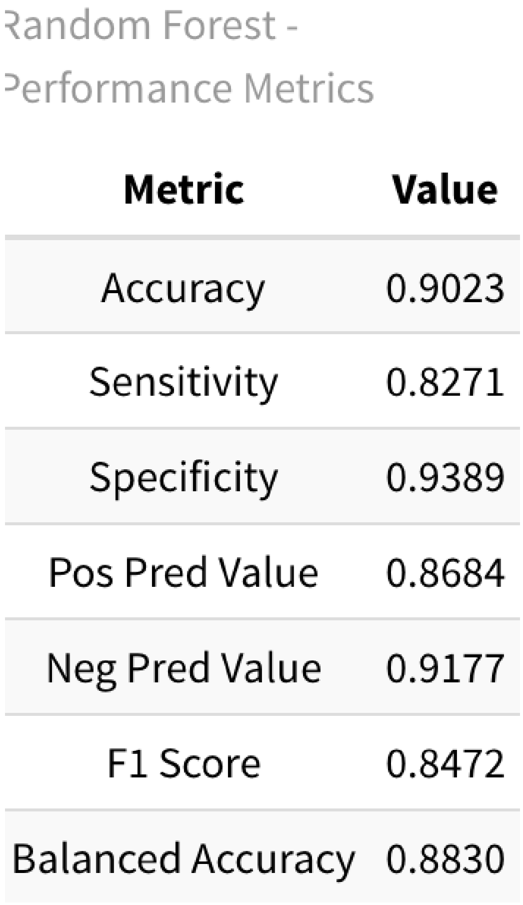
Threshold Tuning F1

- LASSO (λmin.1se) - Best cutoff threshold : 0.38 LASSO
- Random Forest - Best cutoff threshold : 0.42



LASSO (λmin.1se) - Performance Metrics

Metric	Value
Accuracy	0.7847
Sensitivity	0.7749
Specificity	0.7895
Pos Pred Value	0.6420
Neg Pred Value	0.8780
F1 Score	0.7022
Balanced Accuracy	0.7822

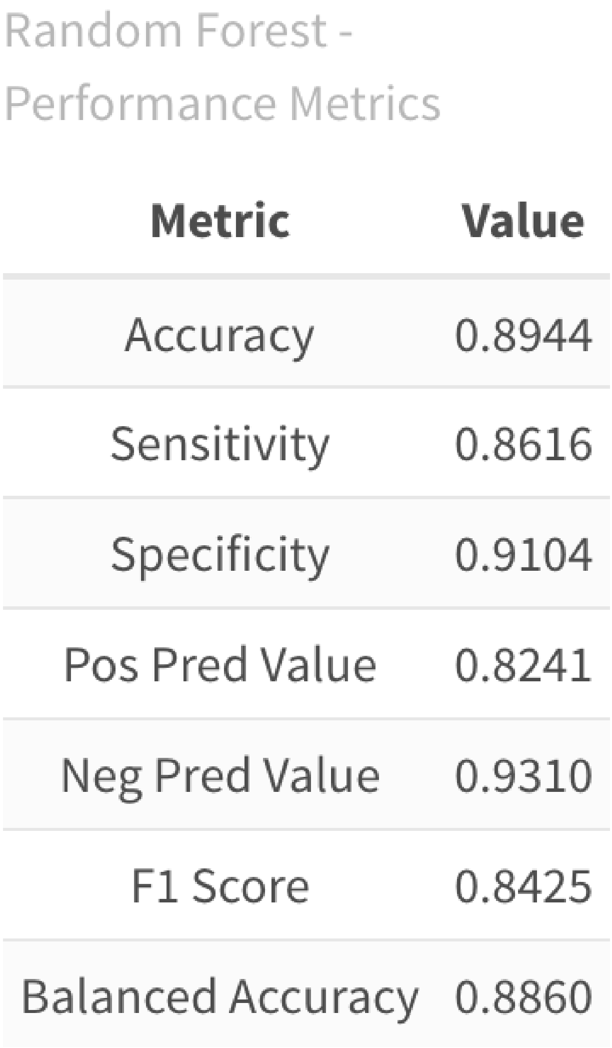


Random Forest - Performance Metrics

Metric	Value
Accuracy	0.9023
Sensitivity	0.8271
Specificity	0.9389
Pos Pred Value	0.8684
Neg Pred Value	0.9177
F1 Score	0.8472
Balanced Accuracy	0.8830

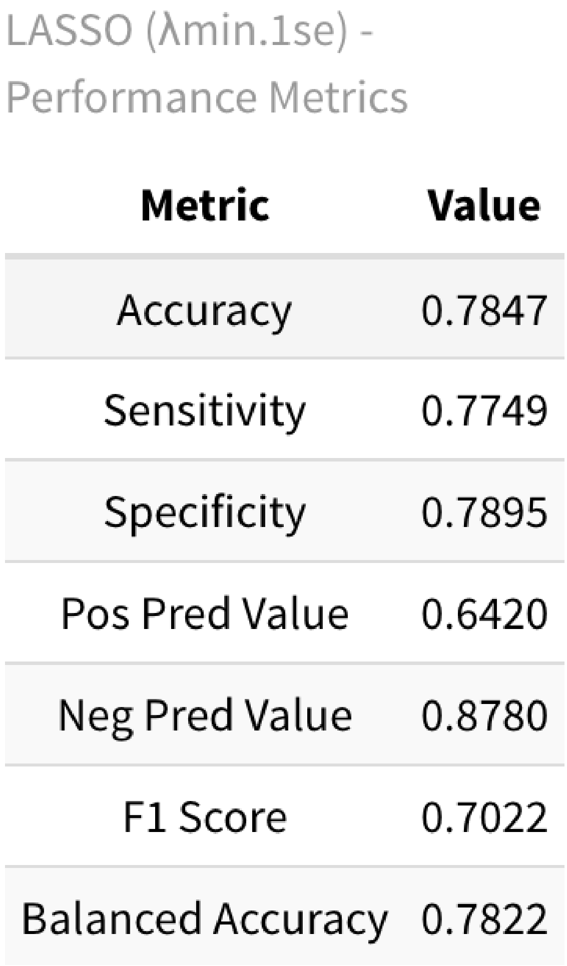
Threshold Tuning ROCR

- LASSO (λmin.1se) - Best cutoff threshold: 0.37
- Random Forest - Best cutoff threshold : 0.38



Random Forest - Performance Metrics

Metric	Value
Accuracy	0.8944
Sensitivity	0.8616
Specificity	0.9104
Pos Pred Value	0.8241
Neg Pred Value	0.9310
F1 Score	0.8425
Balanced Accuracy	0.8860



LASSO (λmin.1se) - Performance Metrics

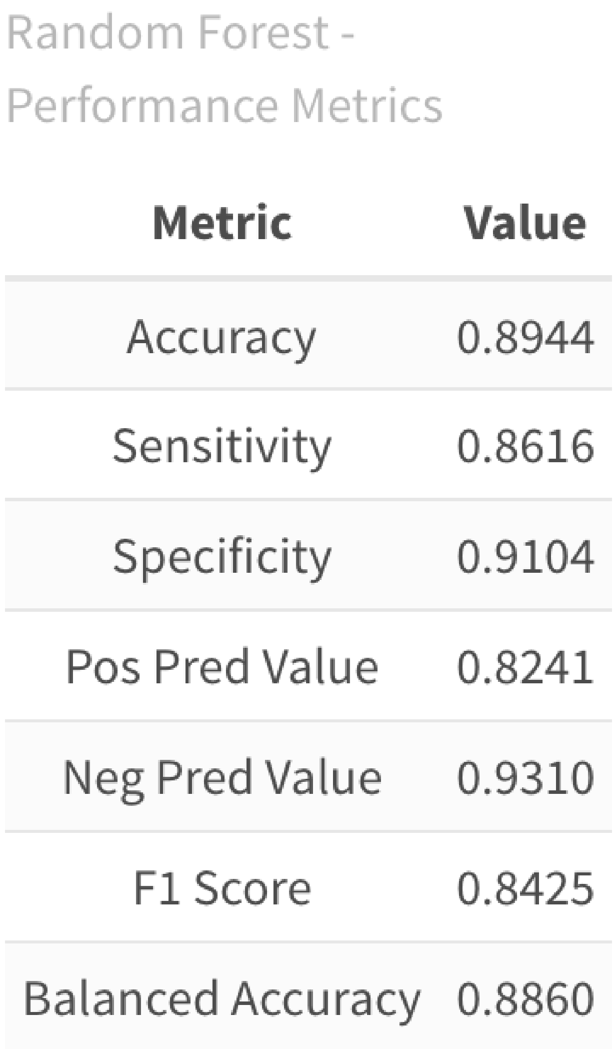
Metric	Value
Accuracy	0.7847
Sensitivity	0.7749
Specificity	0.7895
Pos Pred Value	0.6420
Neg Pred Value	0.8780
F1 Score	0.7022
Balanced Accuracy	0.7822

Threshold Tuning

- **False Negatives:**
  - Occur when a booking that will be canceled is predicted as non-canceled.
  - *Risk:* Unanticipated cancellations leading to unoccupied rooms and lost revenue.
- **False Positives:**
  - Occur when a non-canceled booking is predicted as canceled.
  - *Risk:* Unnecessary interventions or follow-ups that may irritate customers or waste resources.
- **Business Scenarios:**
- **High Cost of Missed Cancellations (False Negatives):**
  - In environments where vacant rooms significantly impact revenue, it's better to err on the side of caution.
  - *Action:* Lower the threshold to increase sensitivity, even if it means accepting more false positives.
- **High Cost of Unnecessary Interventions (False Positives):**
  - When over-alerting can harm customer satisfaction or incur excessive follow-up costs, false positives must be minimized.
  - *Action:* Raise the threshold to boost specificity, even if some cancellations are missed.
- **Threshold Tuning Strategy:**
- Balance sensitivity and specificity using metrics like the F1 score or Youden’s Index.
- Adjust thresholds based on a cost-benefit analysis reflecting the specific operational priorities and customer impact.

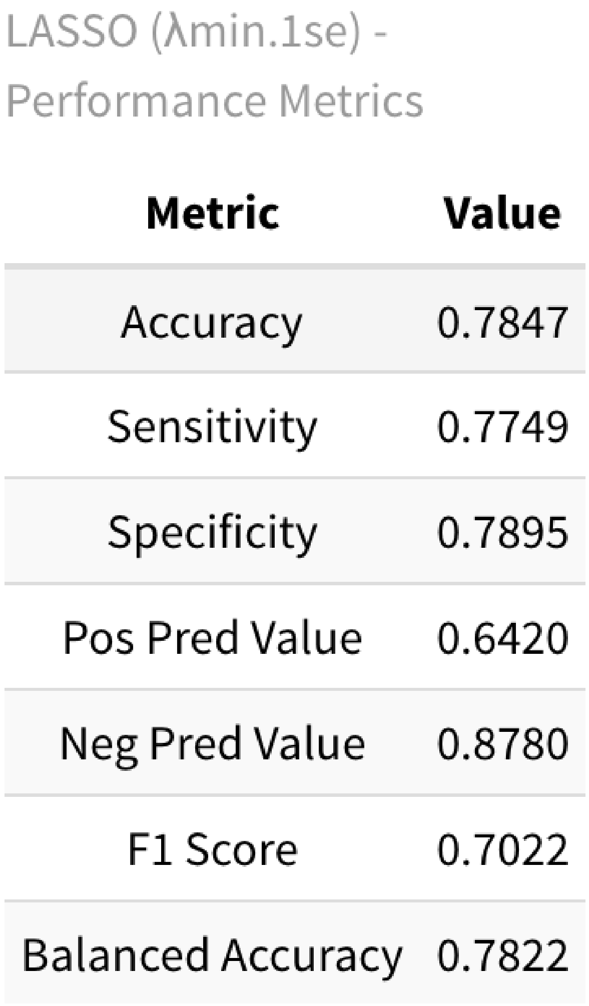
Final Model Comparisons

- **Random Forest Outperforms on Accuracy**
  - Accuracy of ~0.8937 vs. ~0.7938 for LASSO
  - Stronger sensitivity and specificity, leading to higher F1 and balanced accuracy
- **LASSO Excels in Interpretability**
  - Directly shows which features increase or decrease cancellation risk
  - Easier to communicate insights and justify decisions to stakeholders
- **Choosing the Right Model**
  - **If top predictive power is paramount:** Random Forest is the clear winner
  - **If clarity and simplicity matter:** LASSO offers a more transparent view of feature impacts
- **Practical Implications**
  - Business context determines which trade-off is most valuable
  - Threshold tuning can further tailor each model's sensitivity or specificity to operational needs



Random Forest - Performance Metrics

Metric	Value
Accuracy	0.8944
Sensitivity	0.8616
Specificity	0.9104
Pos Pred Value	0.8241
Neg Pred Value	0.9310
F1 Score	0.8425
Balanced Accuracy	0.8860



LASSO (λmin.1se) - Performance Metrics

Metric	Value
Accuracy	0.7847
Sensitivity	0.7749
Specificity	0.7895
Pos Pred Value	0.6420
Neg Pred Value	0.8780
F1 Score	0.7022
Balanced Accuracy	0.7822

Takeaways

1. **Can we predict how likely a booking is to be canceled?**
    1. *Yes*. Our LASSO and Random Forest models demonstrate strong predictive power, accurately identifying which bookings are most at risk of cancellation. The RF model is stronger at making predictions while the LASSO provides more insight to key factors
2. **What factors affect how likely a booking is to be canceled?**
    1. Our analysis highlights several **key drivers** of cancellation risk:
        1. **Lead Time (log-transformed):** Longer lead times correlate with higher cancellation probabilities.
        2. **Average Room Price:** Bookings with higher room prices are more prone to cancellation.
        3. **Special Requests:** Fewer special requests often indicate a higher risk of cancellation.
        4. **Market Segment & Arrival Month:** Certain segments (e.g., online vs. offline) and months exhibit distinct cancellation patterns.

- **Real-World Implications**
- **Proactive Strategies:** Armed with these insights, hotels can adjust pricing, offer incentives, or send targeted reminders to reduce cancellations.
- **Resource Allocation:** Predictive accuracy helps hotels plan staffing and room availability, minimizing revenue loss from vacant rooms.
- **Threshold Tuning:** By balancing false positives and false negatives, businesses can tailor the model’s sensitivity to their unique operational costs and customer service goals.