Hotel Cancellations

Predicting hotel booking cancellations to decrease uncertainty and optimize revenue

Lauma Ustupa | Ironhack

Ever heard of overbooking ?

The Data

Resort Hotel in Algarve, Portugal

Booking transactional data

Period ***Jul 2015 - Aug 2017***

Shape (39 665, 32)

Main features:

Guest:

- Country
- Market Segment/ Distribution Channel/ Customer Type
- Repeated Guest/ Previous Cancellations/ Not Cancelled

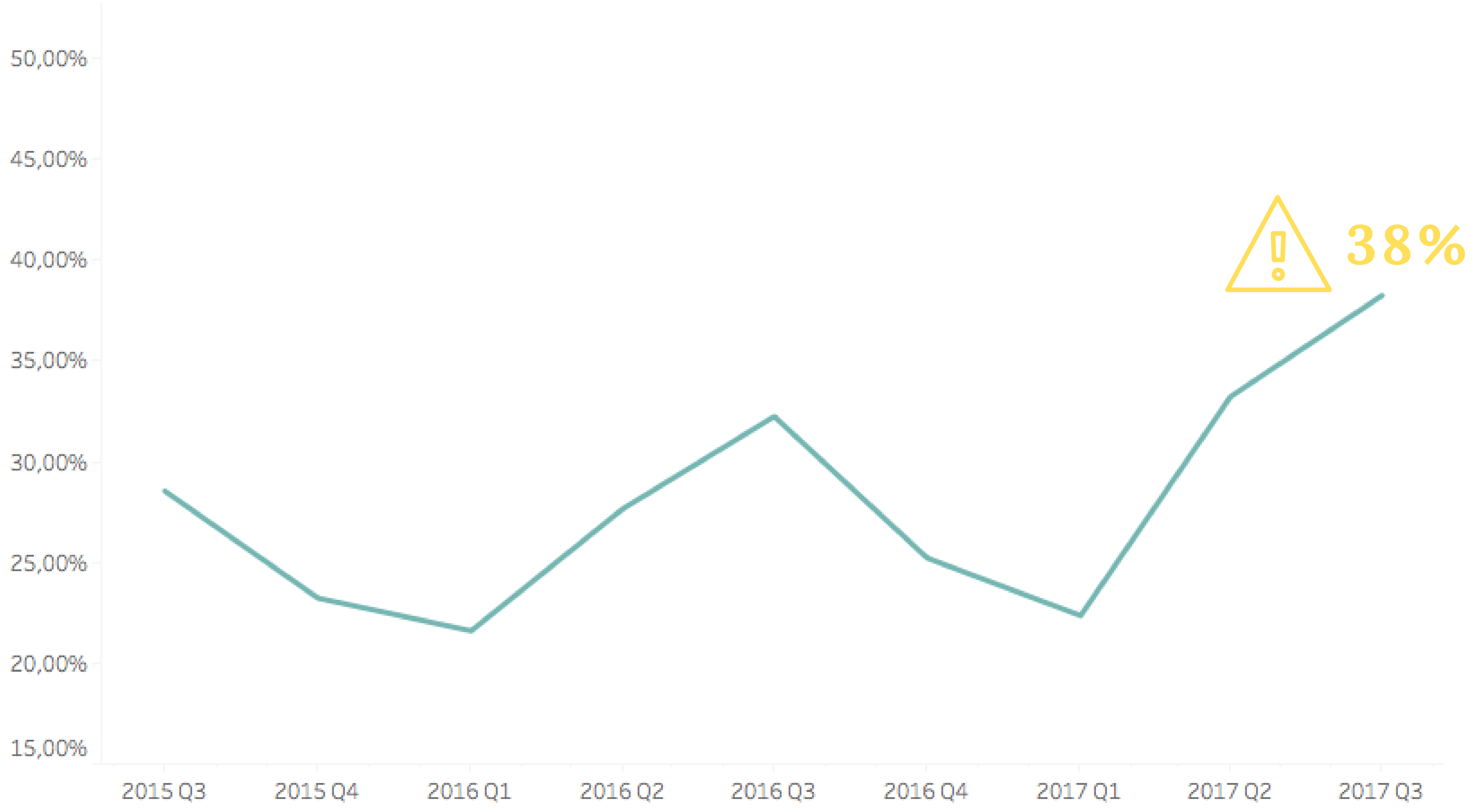
Booking:

- Arrival Date
- Stay in Week/ Weekend Nights
- Number of Adults/ Children/ Infants
- Booked & Assigned Room Type
- Number of Booking Changes/ Days in Waiting List
- Average Daily Rate/ Deposit Type
- Lead Time (days since booking was made)
- Services: Meal, Parking, Special requests

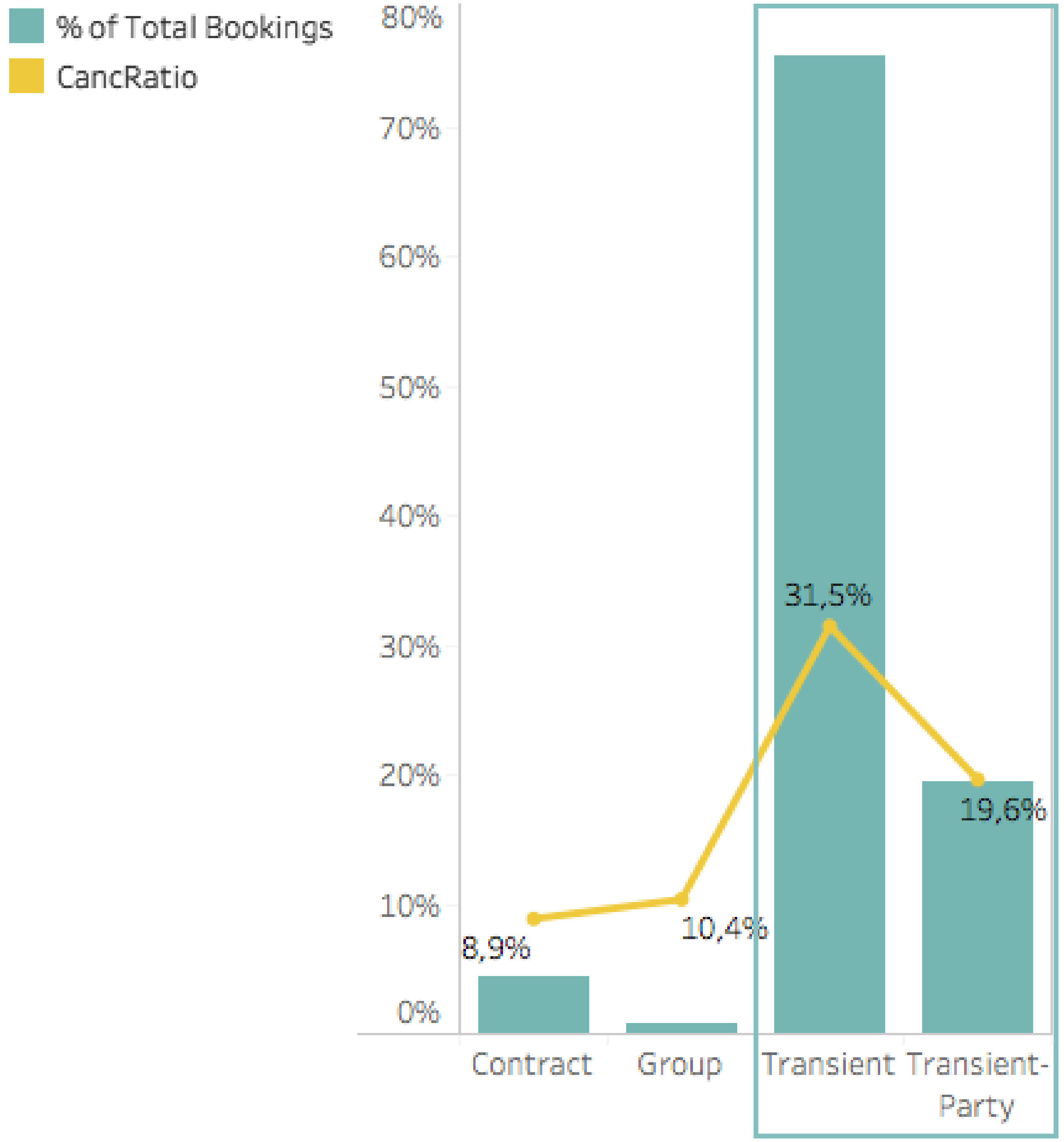
Target:

- Is Cancelled: Yes/No

Cancellation Ratio Over Time



Bookings vs Cancellation Ratio per Customer Type



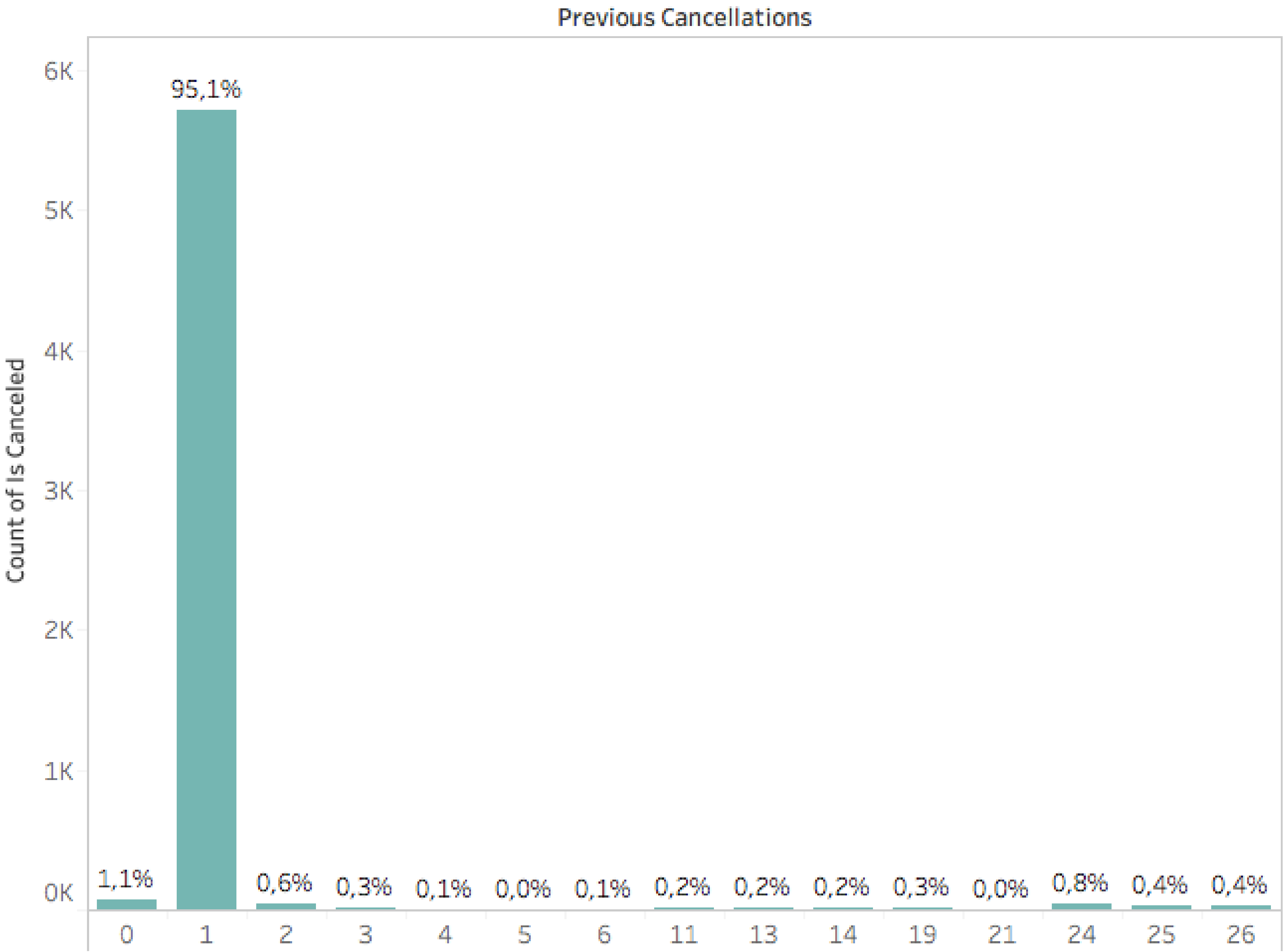
95%

of hotel bookings come from transient segment (individual, non-group)

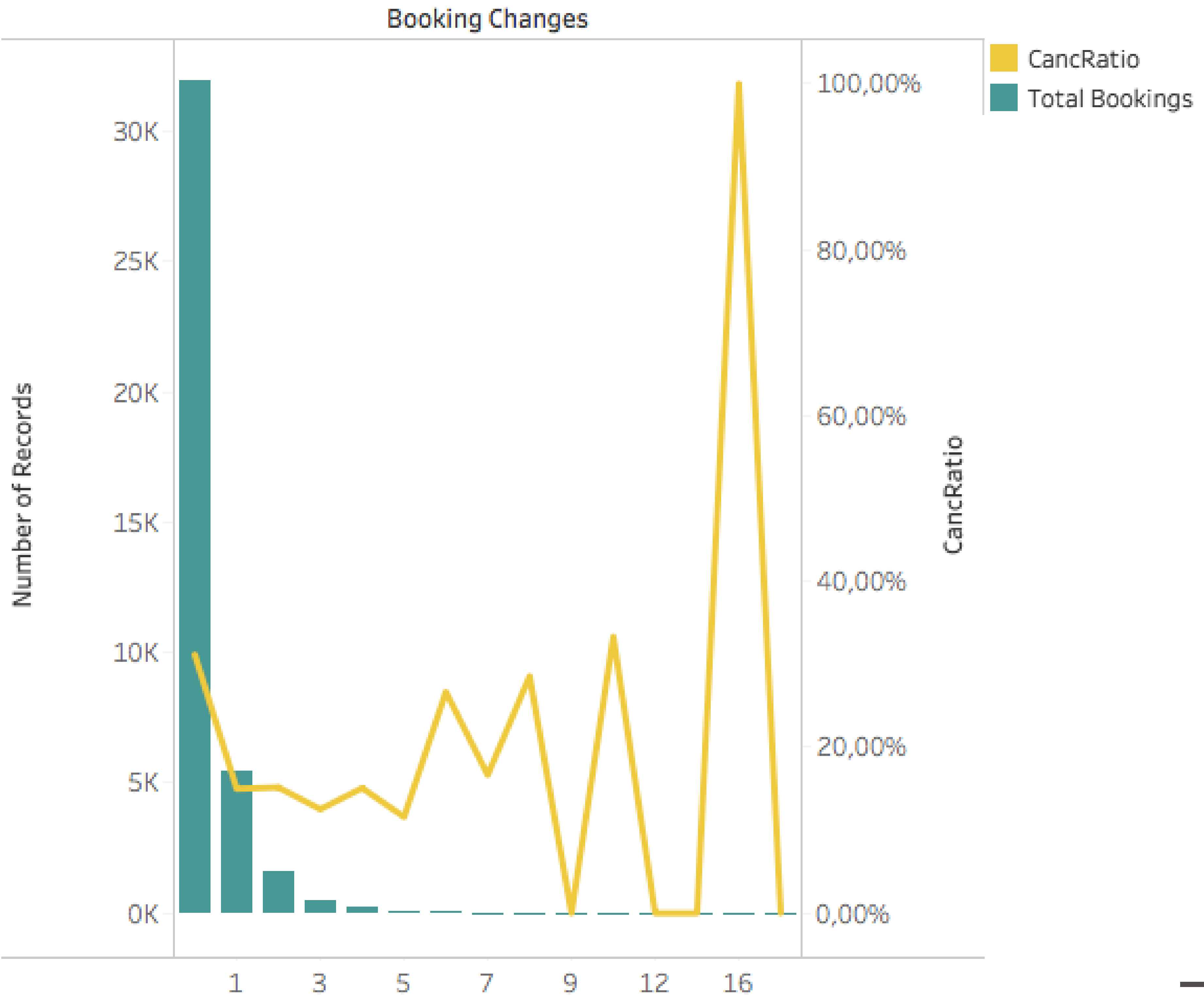
51%

of these bookings are being cancelled

Cancellation Ratio vs Previous Cancellations



Booking Changes vs Cancellation Ratio



Building ML Model

Prediction of bookings likely to be cancelled

Steps

- *Feature selection & feature engineering, OHE*
- *Selecting best performing models*
- *Tuning hyper-parameters*
- *Result Analysis*

Model Selection

- *Selecting best performing models (using Cross Validation with Kfold)*

Model	Accuracy	Precision
Logistic Regression	0,775	0,707
Gausian Naive Bayes	0,568	0,373
Decision Tree	0,760	0,566
SVM	0,722	0,000
Random Forest	0,813	0,727
Gradient Boosting Classifier	0,806	0,734
XG Boost	0,806	0,733

Minimize the number of false positives

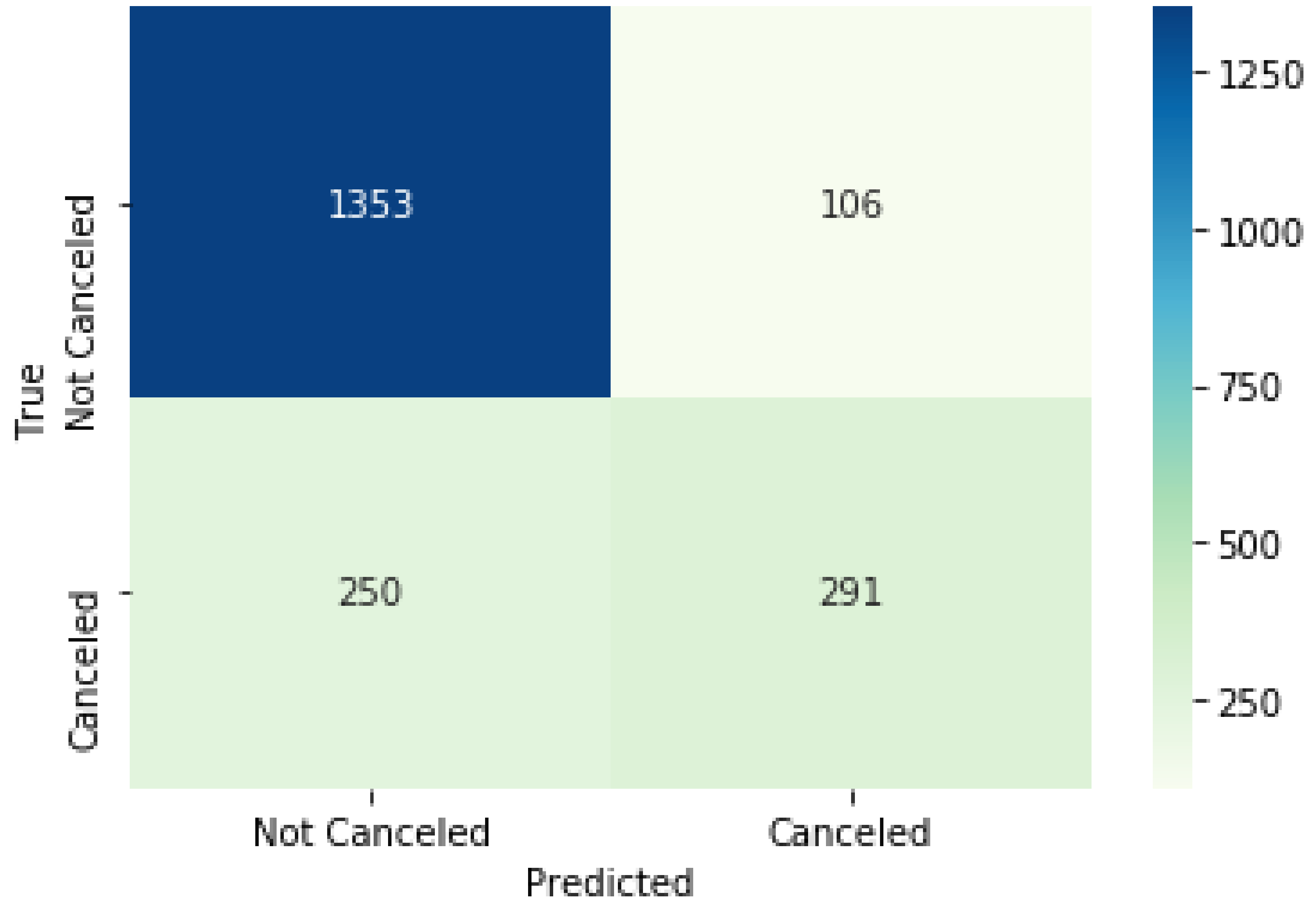
Tuning Hyperparameters

- *Using Grid Search Cross Validation (CV = 5)*
- *Best performing parameters per model:*

	Random Forest	Gradient Boosting	XG Boost
	N estimators: 50 Max Depth: 15 Max Features: 10 Class Weight: balanced	N estimators: 100 Max Depth: 15 Max Features: 10	N estimators: 100 Max Depth: 15 Max Features: 5 Scale pos Weight: 2.59
	Accuracy: 0.802	Accuracy: 0.812	Accuracy: 0.793
Test data	Accuracy: 0.806 Precision: 0.624	Accuracy: 0.822 Precision: 0.732	Accuracy: 0.809 Precision: 0.638

Result Analysis: Confusion Matrix

Gradient Boosting Classifier

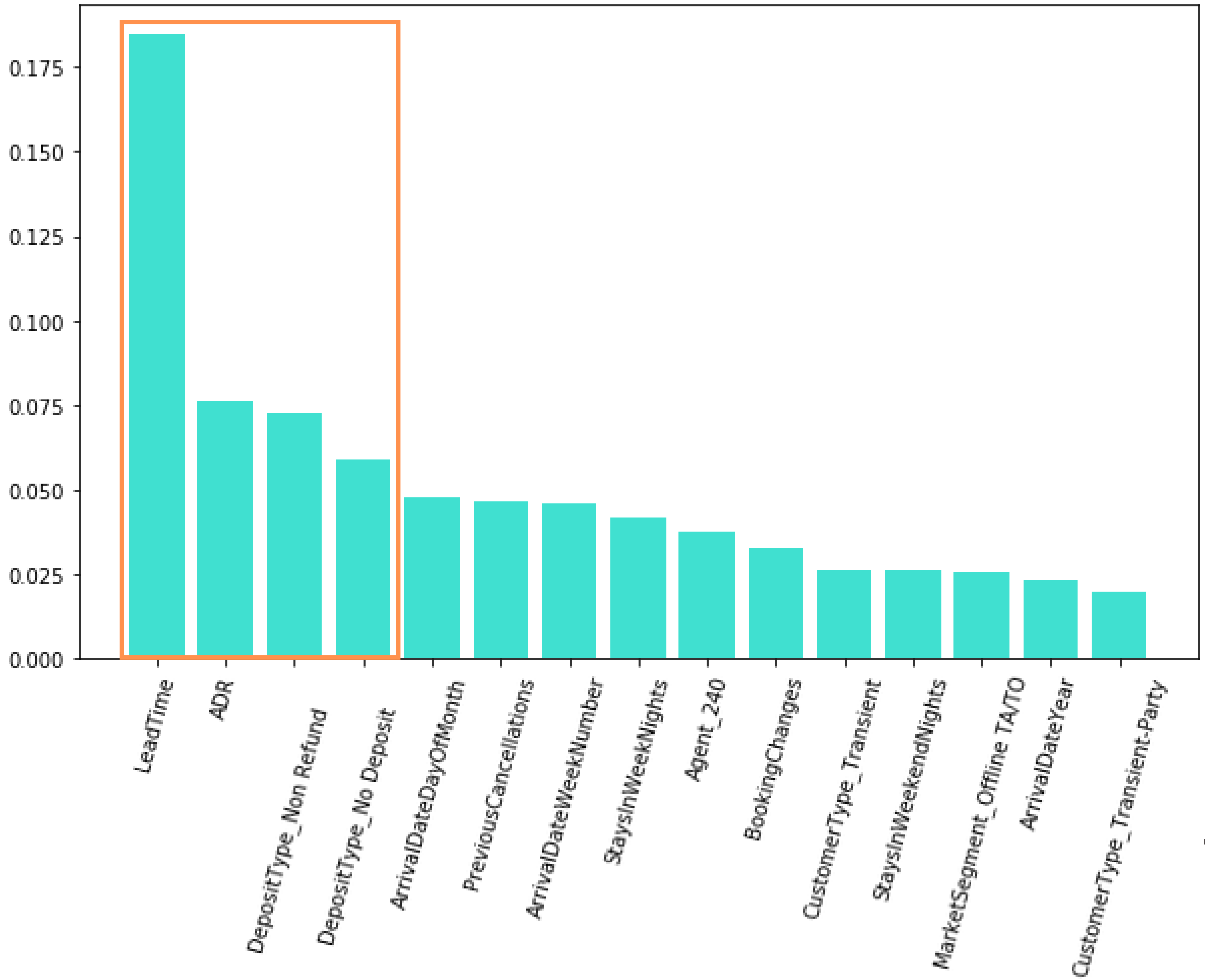


False Positives

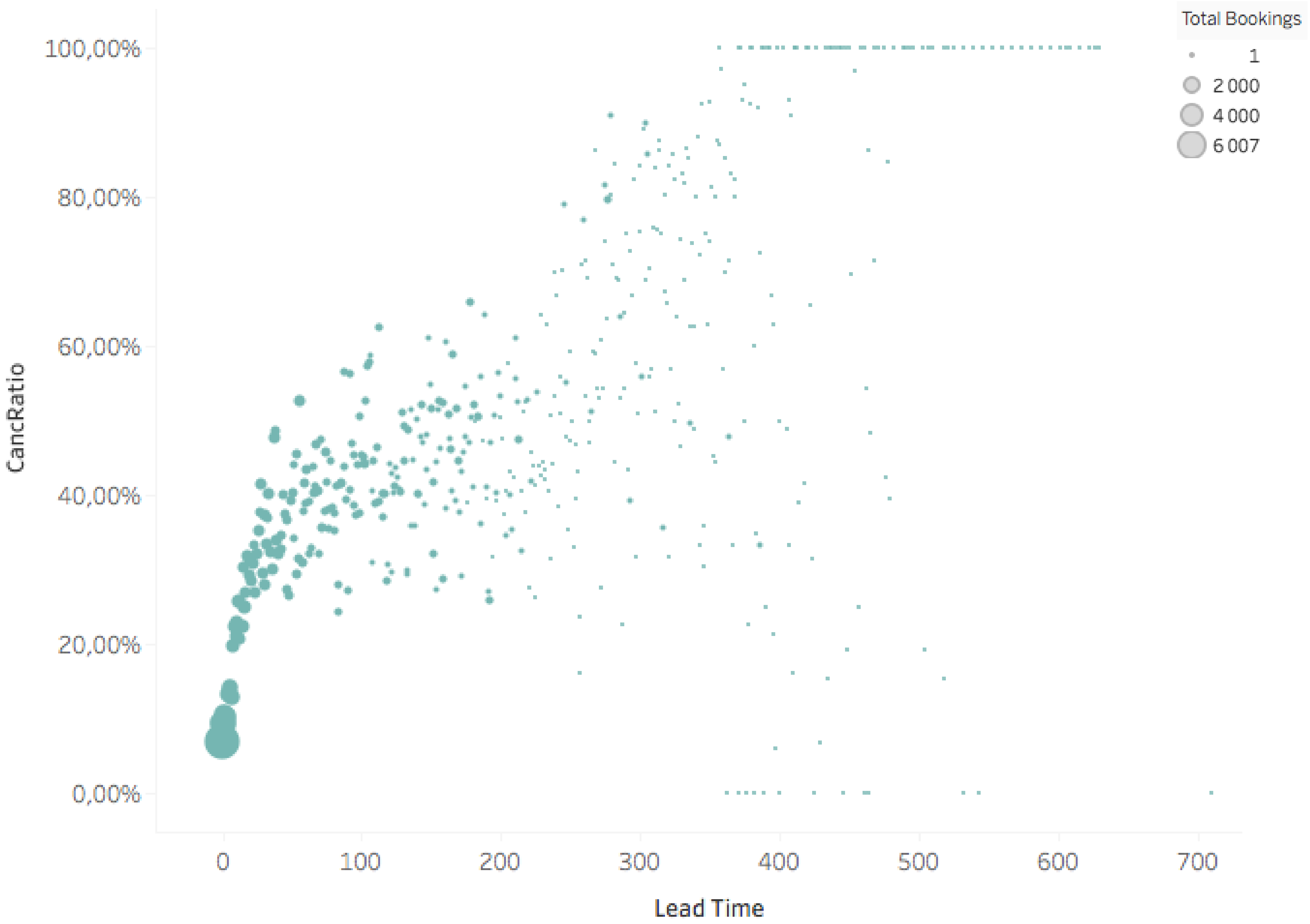
Result Analysis: Feature Importances

Gradient Boosting Classifier

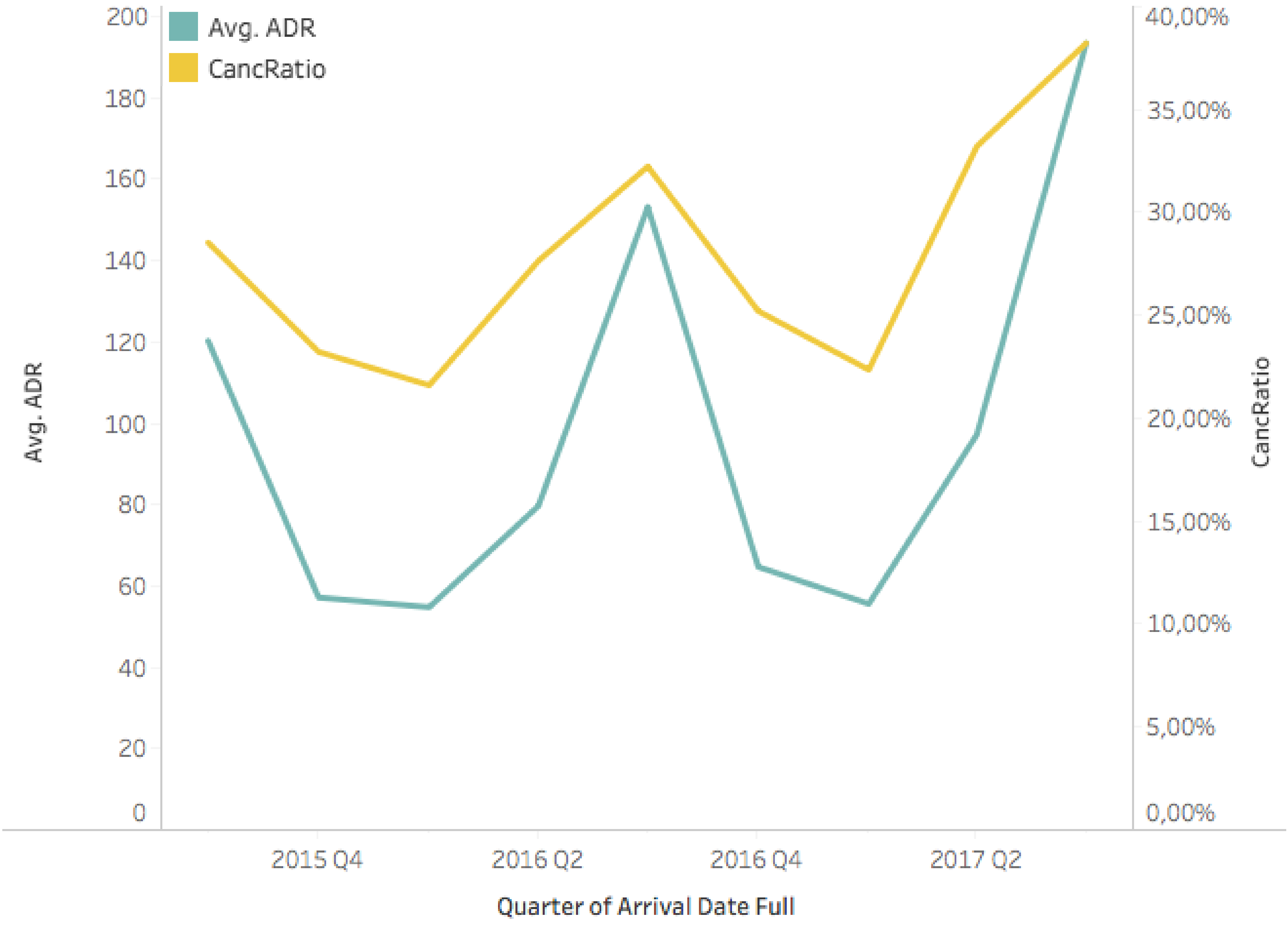
Lead Time
ADR
Deposit Type



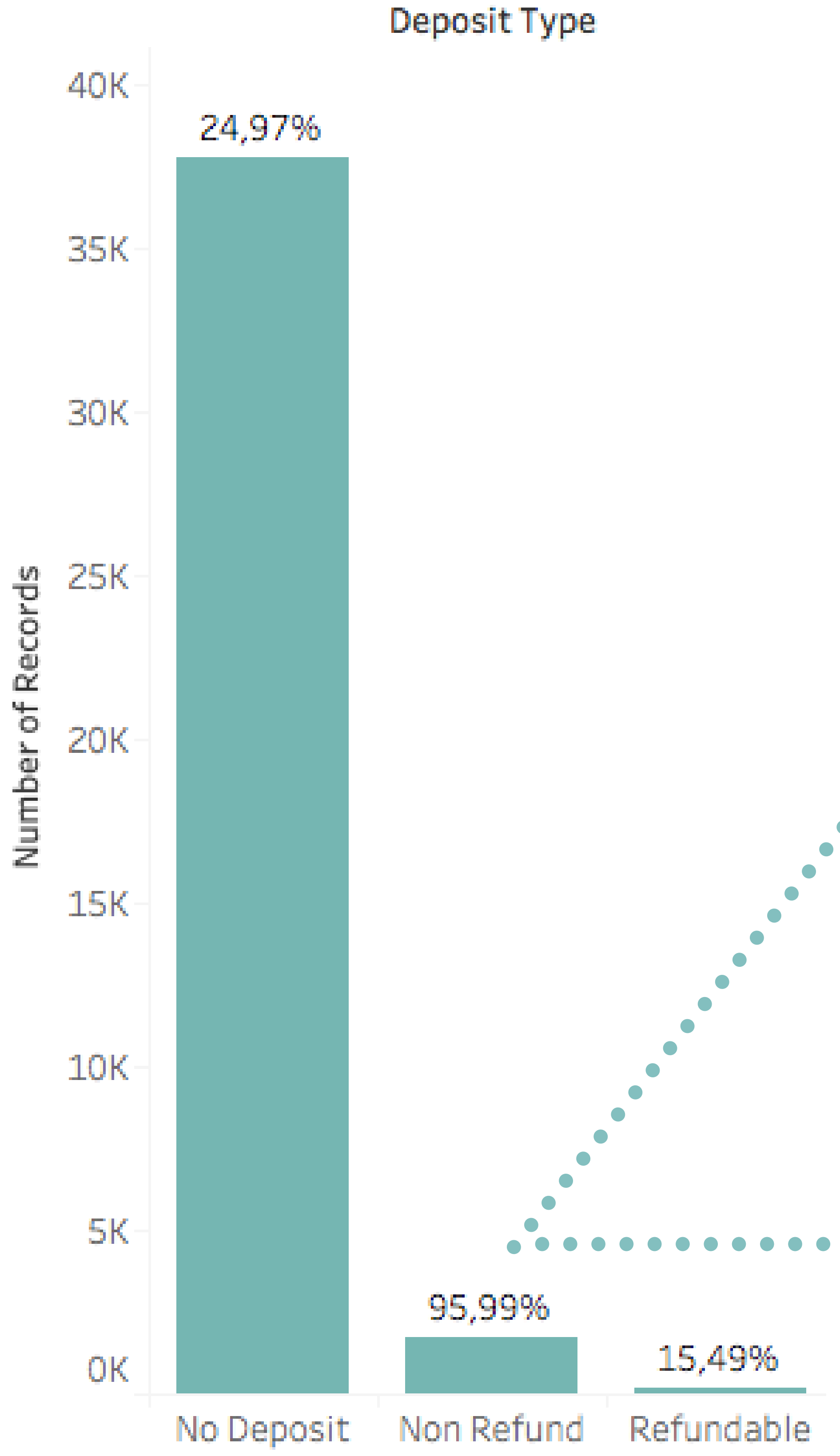
Cancellation Ratio vs Lead Time

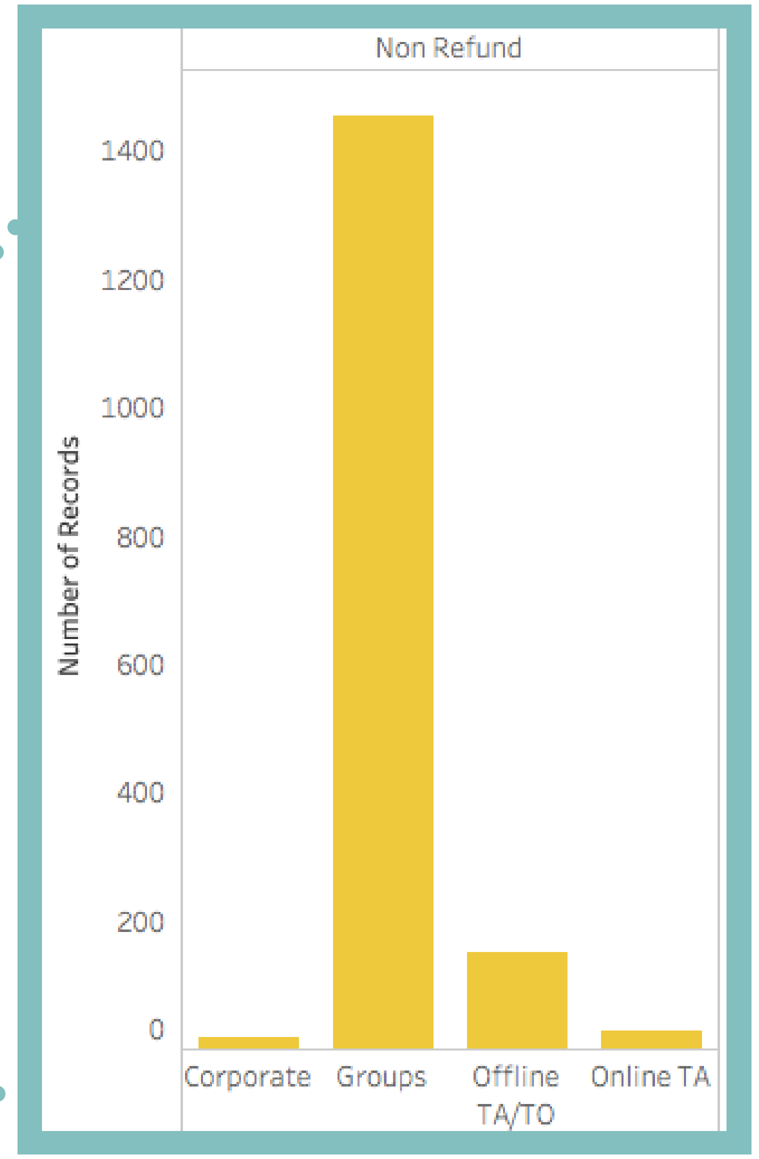


Cancellation Ratio vs ADR



Cancellation Ratio vs Deposit Type





Conclusion & further exploration

- *Better decisions on bookings to accept or reject (if bookings on request/in waitlist)*
- *Calculate* **Hotel Net Demand** = *Total demand - Bookings likely to be cancelled*
    - *Indication of how many rooms to* ***oversell*** *(use probability to identify bookings most likely to be cancelled)*
- *Continuous* **Model Improvement**: *daily reports on booking status to evaluate the predictions and adjust the model based on results*
- *Build a model to predict* ***cancellation ratio per day*** *by aggregating booking data per each day of the year*

Thank you!