



# The Hotel Dilemma

April, Monique, Nick, and Tony

# WHAT BROUGHT US HERE TODAY?

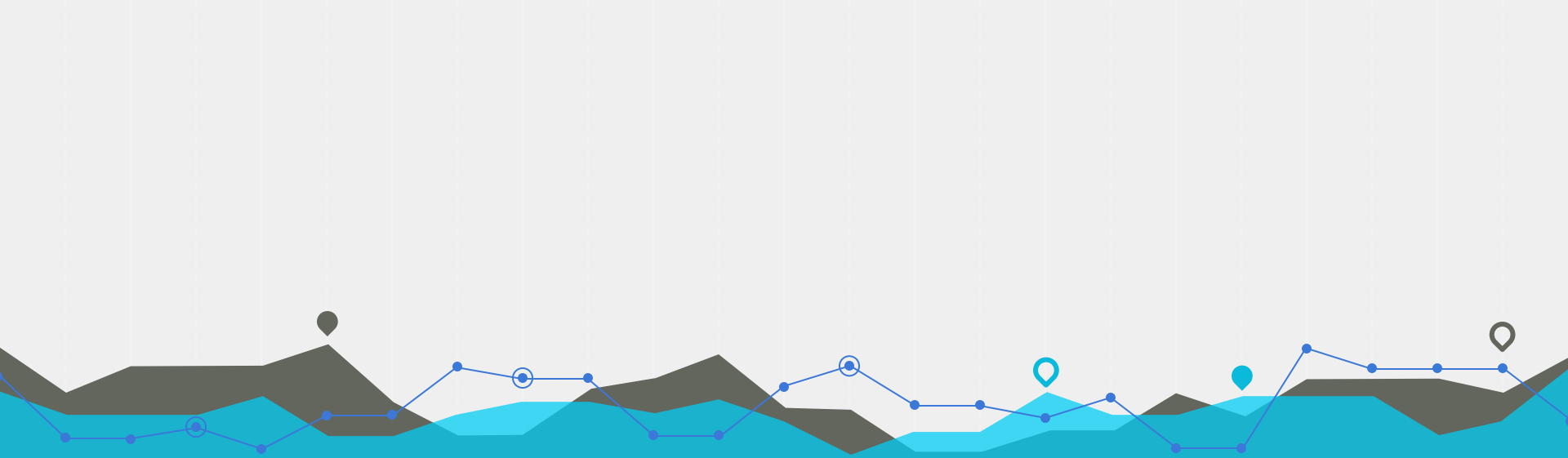
## RESEARCH QUESTIONS

1. Using data and ML models, can hotel reservation cancellations be predicted?
2. If yes to the above, which models and methods most accurately predict hotel reservation cancellations?

## OBJECTIVES

1. Build a ML model that can predict whether a hotel reservation will be cancelled.
2. Analyze and understand data via organization, visualization, and dashboards.



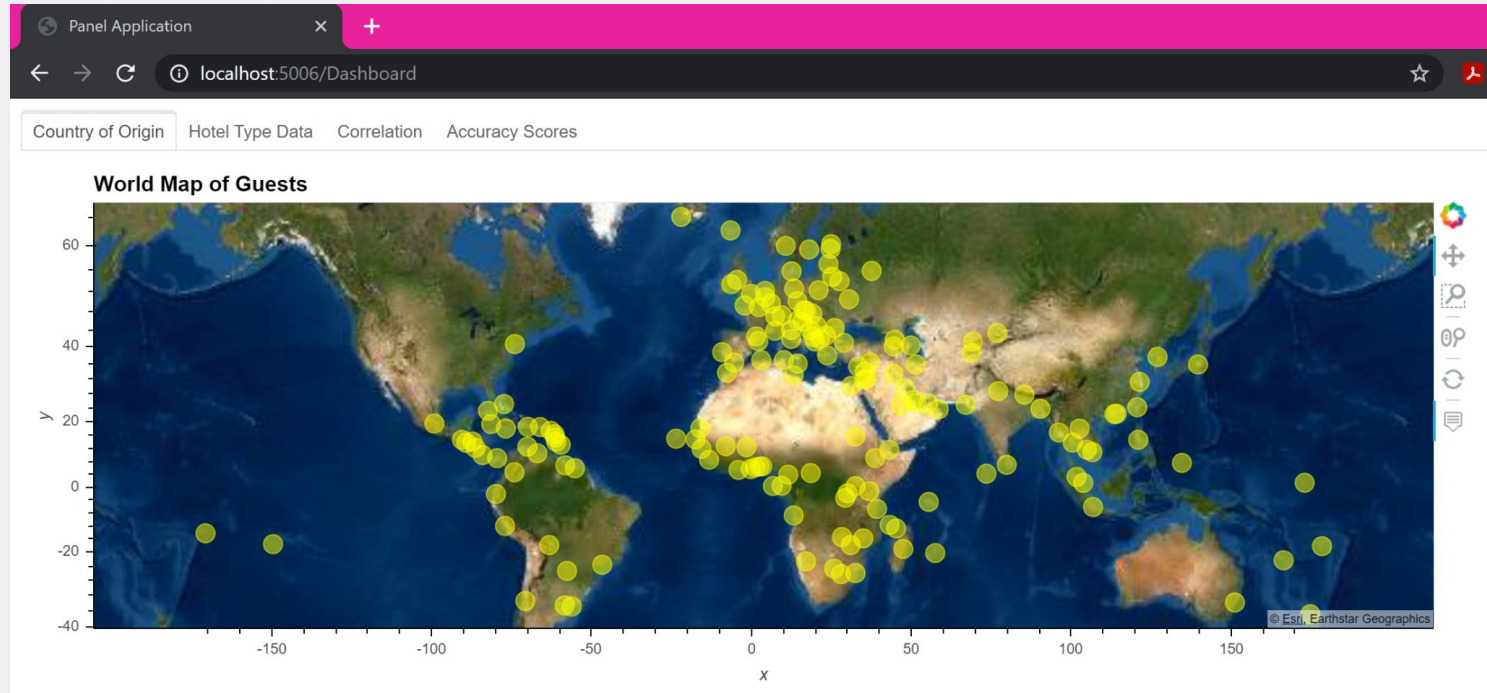


# DATA VISUALIZATION

# Hotels in Portugal - Guests Data

(July 2015 through August 2017)

## Country of origin



# Hotels in Portugal - Guests Data

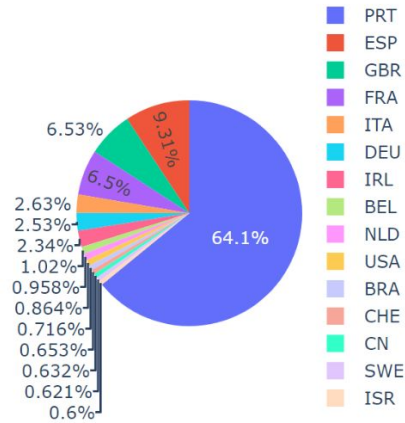
(July 2015 through August 2017)

## Country of origin

Percentage of Total Guest Reservations by Year and Country

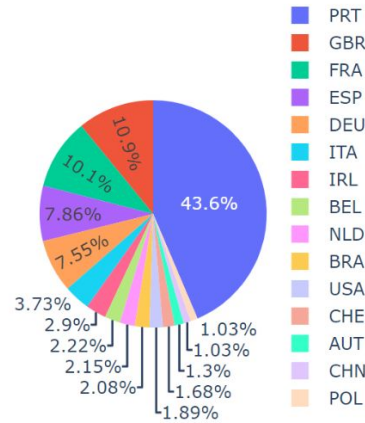
2015

Top 15 Countries



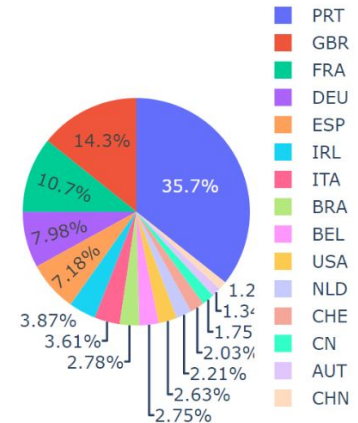
2016

Top 15 Countries



2017

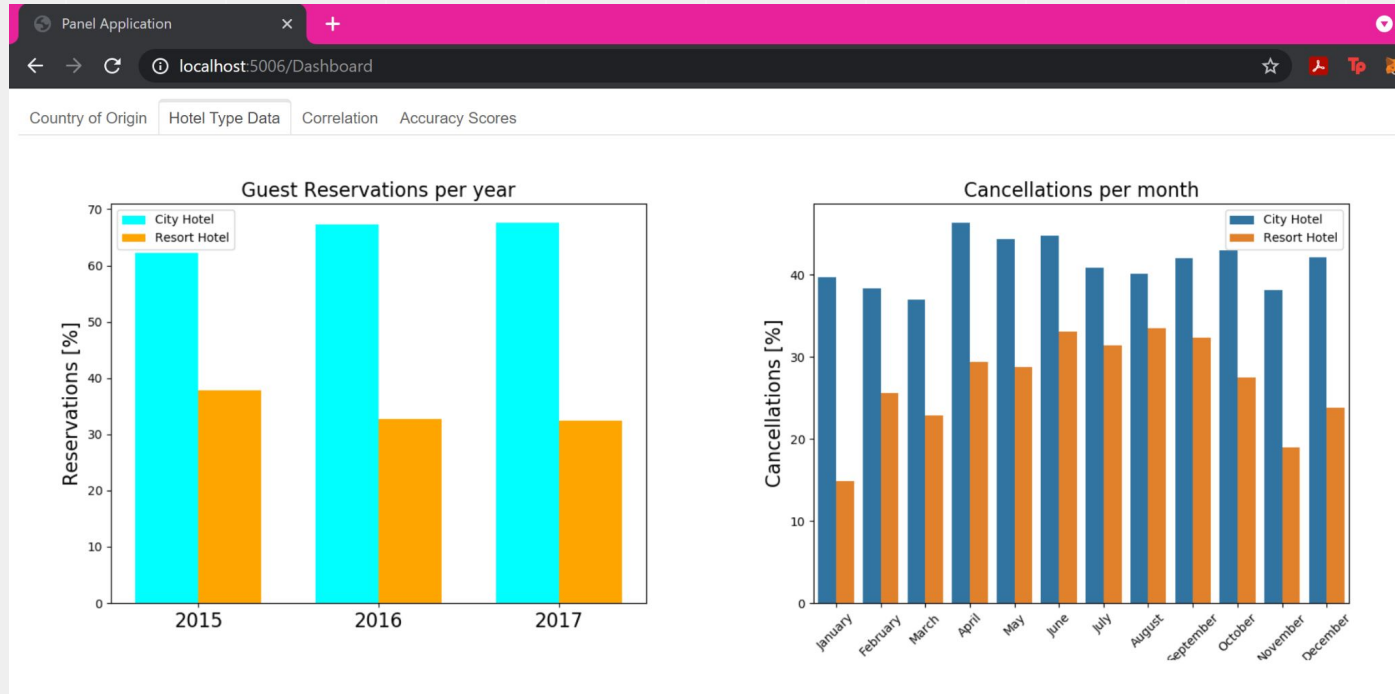
Top 15 Countries

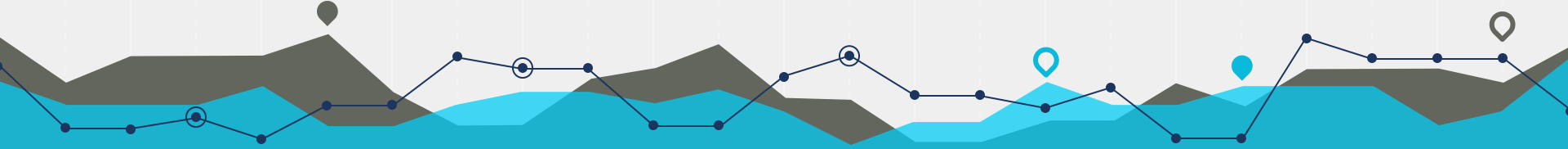
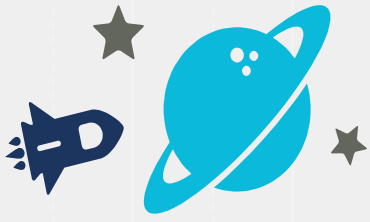


# Hotels in Portugal - Guests Data

(July 2015 through August 2017)

## City Hotel vs. Resort Hotel





# Data Cleaning & Preparation

# Data: The Source?

- Data from an article written by Nuno Antonio, Ana Almeida, and Luis Nunes, for Data in Brief, Volume 22, February 2019 (ScienceDirect Journal)
- Data was downloaded and cleaned by Thomas mock and Antoine Bichat on February 11th, 2020 (we had to conduct additional cleaning)
- Original source unknown, appears to be extracted from a site such as Expedia





# The Basics

- 31 Columns (or “features”)
- 119,391 Rows, each an instance of a hotel reservation (very robust)
- Looks at bookings data between July 2015 and August 2017
- All data is from Portugal, and includes both a city hotel and a resort hotel
- Example of feature descriptions in the academic text:

Variable	Type	Description	Source/Engineering
<i>ADR</i>	Numeric	Average Daily Rate as defined by [5]	BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights
<i>Adults</i>	Integer	Number of adults	BO and BL
<i>Agent</i>	Categorical	ID of the travel agency that made the booking <sup>a</sup>	BO and BL
<i>ArrivalDateDayOfMonth</i>	Integer	Day of the month of the arrival date	BO and BL

## Our Process

- Import CSV
- Check nulls and assign values if needed
- Check data types (address non-integer types)
- Utilize labelencoder on “object” data types
- Run a correlation matrix
- Delete unneeded features
- Final data set for modelling ⇒

Data columns (total 29 columns):

#	Column	Non-Null Count	Dtype
0	hotel	119390 non-null	int64
1	is_canceled	119390 non-null	int64
2	lead_time	119390 non-null	int64
3	arrival_date_year	119390 non-null	int64
4	arrival_date_month	119390 non-null	int64
5	arrival_date_week_number	119390 non-null	int64
6	arrival_date_day_of_month	119390 non-null	int64
7	stays_in_weekend_nights	119390 non-null	int64
8	stays_in_week_nights	119390 non-null	int64
9	adults	119390 non-null	int64
10	children	119390 non-null	int64
11	babies	119390 non-null	int64
12	meal	119390 non-null	int64
13	country	119390 non-null	int64
14	market_segment	119390 non-null	int64
15	distribution_channel	119390 non-null	int64
16	is_repeated_guest	119390 non-null	int64
17	previous_cancellations	119390 non-null	int64
18	previous_bookings_not_canceled	119390 non-null	int64
19	reserved_room_type	119390 non-null	int64
20	assigned_room_type	119390 non-null	int64
21	booking_changes	119390 non-null	int64
22	deposit_type	119390 non-null	int64
23	agent	119390 non-null	int64
24	days_in_waiting_list	119390 non-null	int64
25	customer_type	119390 non-null	int64
26	adr	119390 non-null	int64
27	required_car_parking_spaces	119390 non-null	int64
28	total_of_special_requests	119390 non-null	int64

dtypes: int64(29)

# Preprocessing & Model Preparation

- X = independent variables, or our “features”
- y = dependent variable, or our “was the reservation cancelled?” variable (75k cancelled to 44k not cancelled observations)
- Split into training & testing segments, and scaled
- Tested 5 models:
  - BalancedRandomForest;
  - LogisticRegression; EasyEnsemble
  - Applied naive random oversampling and SMOTEENN to BalancedRandomForest

```
# Define X (independent) and y (dependent) values for modelling, here we se
X = hotel.drop(columns="is_cancelled")
y = hotel["is_cancelled"]
```

```
# Check target counts: canceled (1) or not (0)
y.value_counts()
```

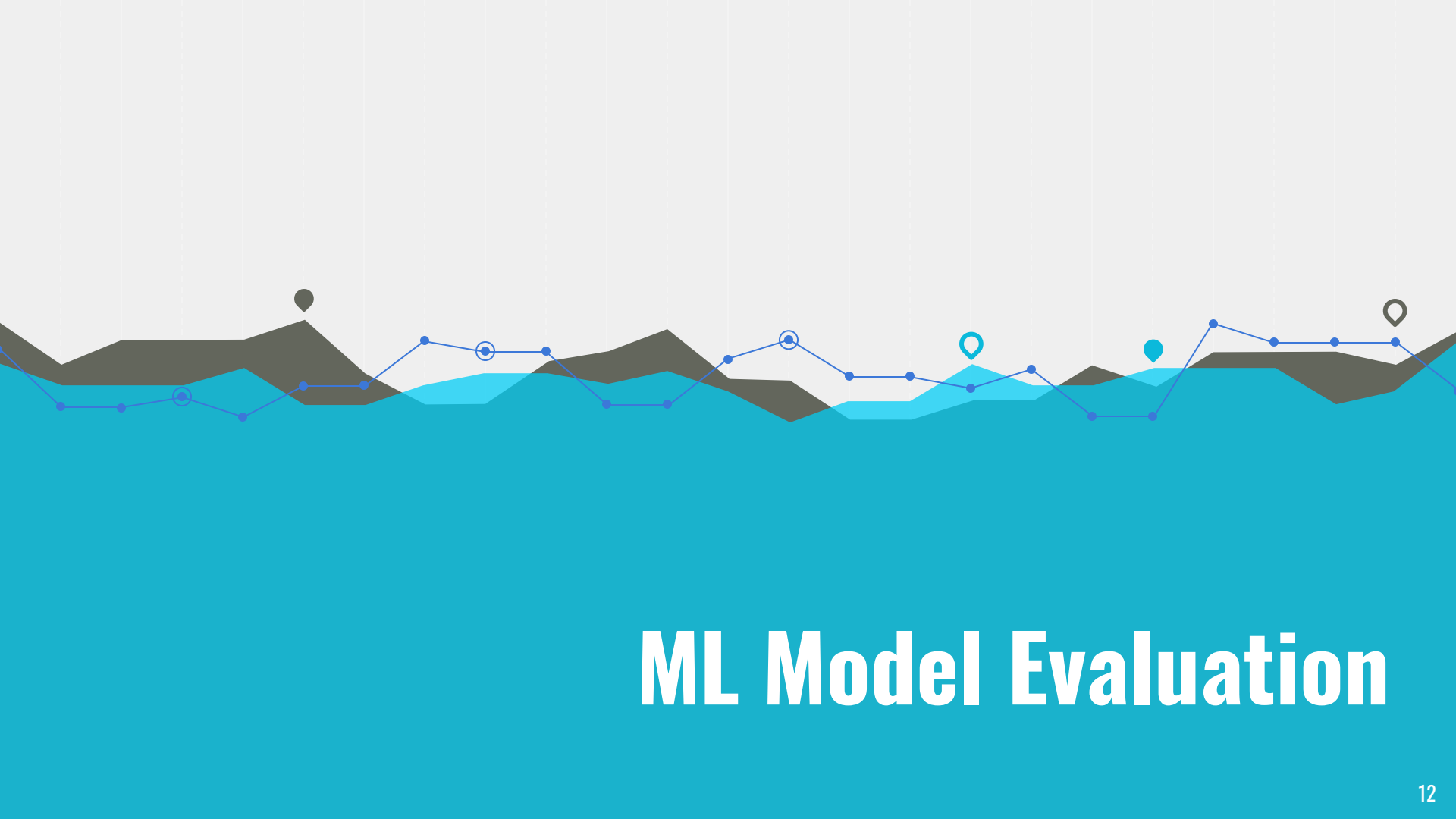
```
0    75166
1    44224
Name: is_cancelled, dtype: int64
```

```
# Split the X and y into X_train, X_test, y_train, y_test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

```
# Create the StandardScaler instance, then scale the X data
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
```

```
# Fit the Standard Scaler with the training data
X_scaler = scaler.fit(X_train)
```

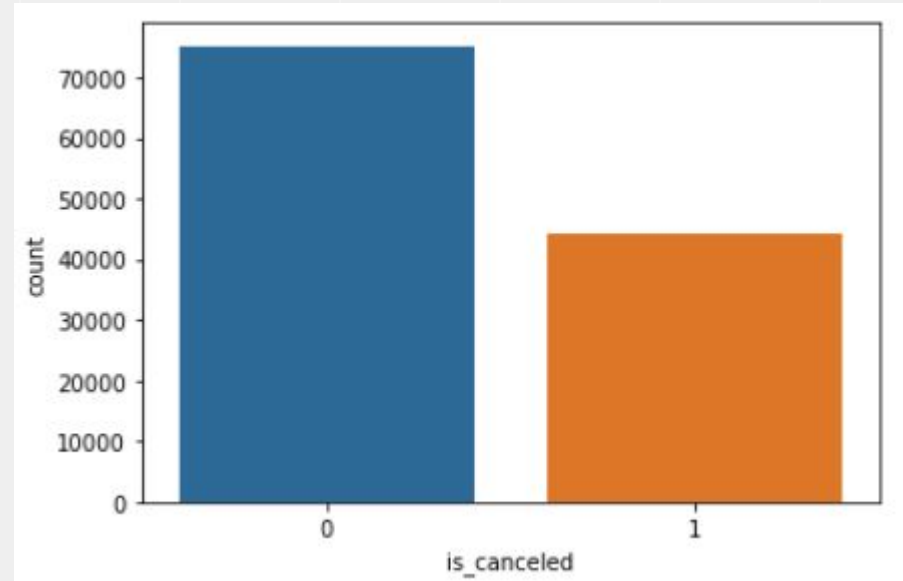
```
# Scale the training and testing data
X_train_scaled = X_scaler.transform(X_train)
X_test_scaled = X_scaler.transform(X_test)
```



# ML Model Evaluation

# Target Variable

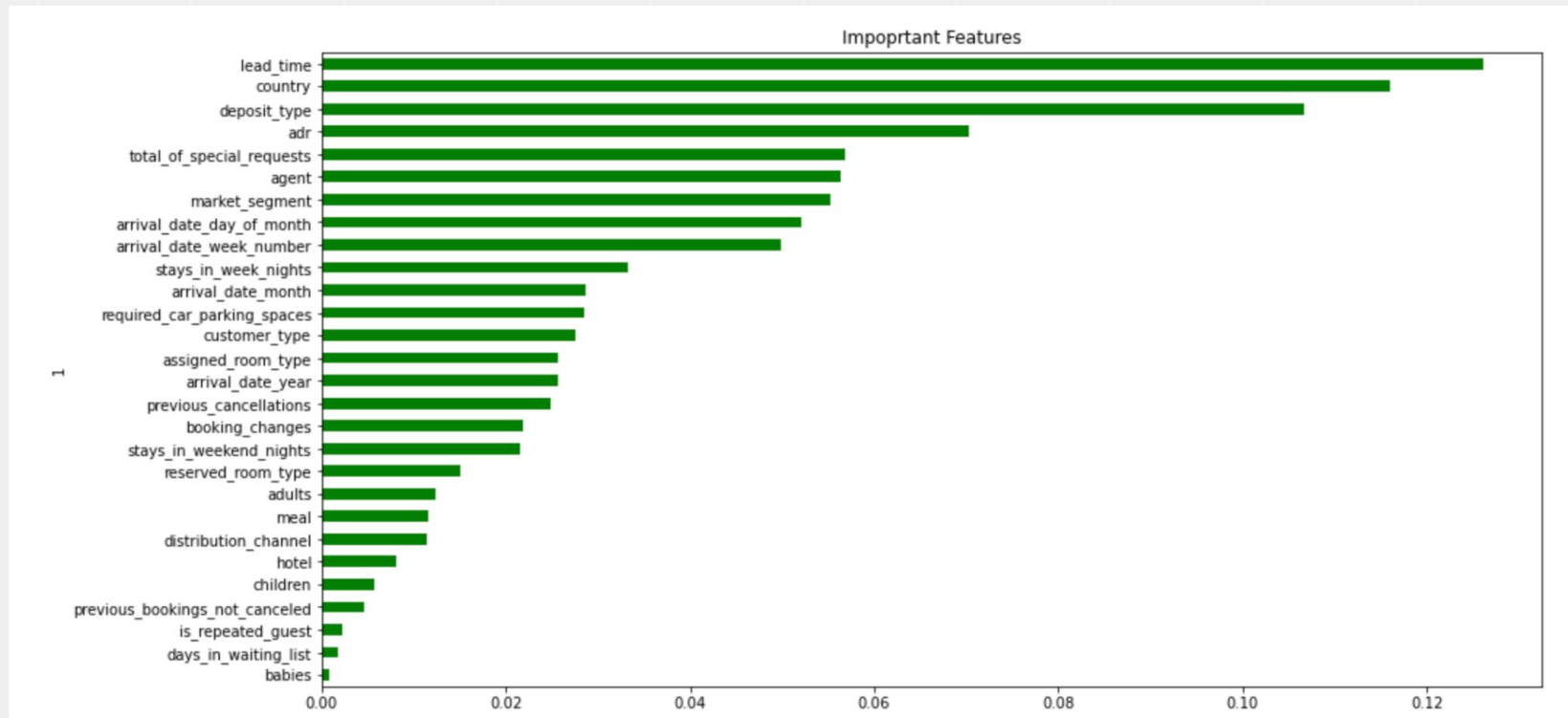
- Reservation canceled (1) or not (0)
- Data is imbalance hence ensemble learning models were chosen



# Model Outcomes

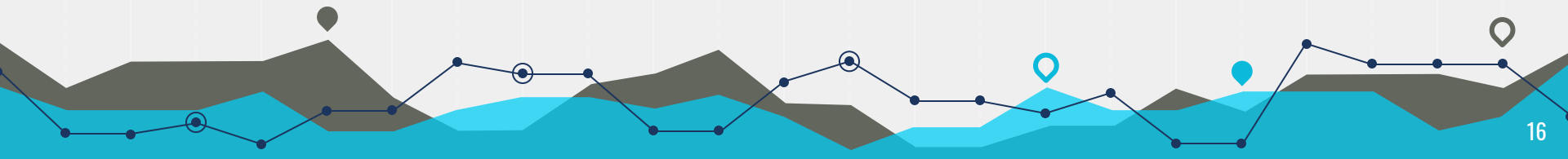
ML Model	Accuracy Score	Avg Precision Score	Avg Recall Score
Balanced Random Forest Classifier	0.89	0.89	0.89
Logistic Regression	0.80	0.80	0.80
Easy Ensemble Classifier	0.82	0.83	0.83
Naive Random Oversampler + BRF	0.88	0.89	0.89
SMOTEENN + BRF	0.90	0.91	0.90

# Feature Importance



# Summary

- ◉ Customers made more city hotel reservations between 2015-2017 than resort hotels.
- ◉ Lead time, country of origin, average daily rate and no deposit/non-refundable were the top features if a booking will be canceled or not.
- ◉ SMOTEENN applied to BalancedRandomForest had the best ML model outcomes to predict hotel reservation cancelations.
  - Aid in data driven decisions to plan supply + demand and personnel coverage as needed.





# THANKS!

**Any questions?**

