

Optimizing User Experience in Hotels Searches by Accurate Price Forecasts

Location: Needham, MA



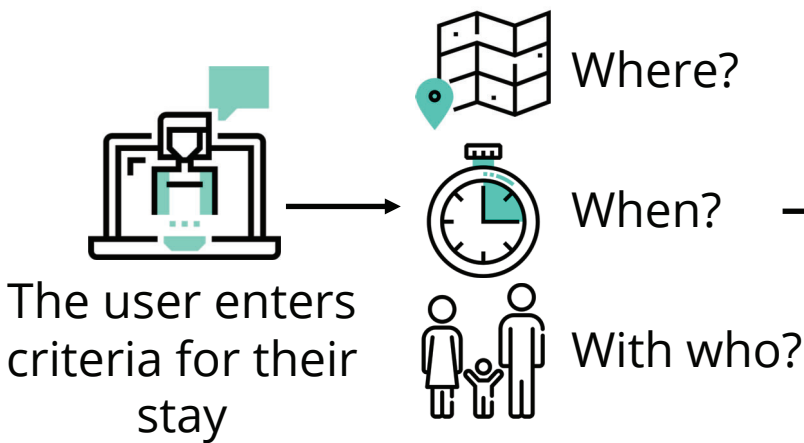
TripAdvisor team: Craig Schmidt, Jeff Palmucci
MIT Advisor & Mentor: Prof. Robert Freund, Renbo Zhao



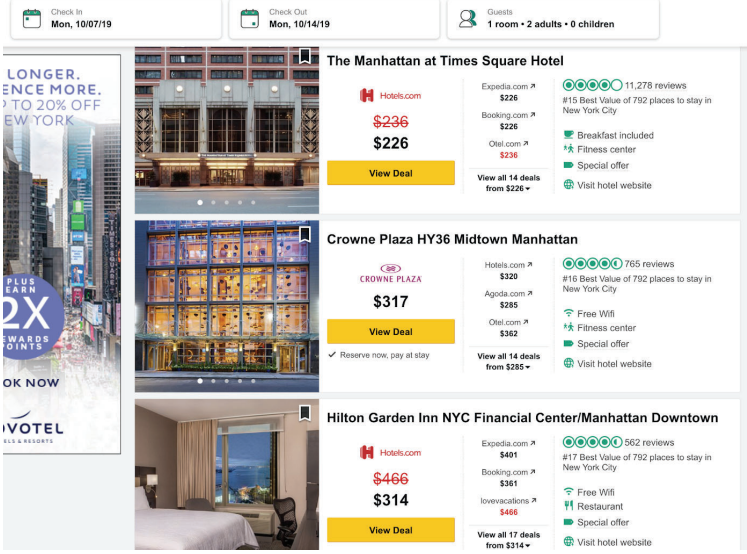
Problem Statement

Getting real-time prices is time-consuming and expensive:
How to avoid requesting prices in advance from 100+ external providers for all 365 possible travel dates and 400+ hotels?
➔ Accurate price forecasts

How does hotels’ page work on TripAdvisor?



Real-time prices are received from external providers (Booking.com, Priceline, ...) via HAC (Hotel Availability Check) calls



Scope: New York City Hotels

Timeline

February - March

March - July

July - August

Exploratory Data Analysis

Iterations of Exploratory Data Analysis
Outlier Detection
Feature Engineering
Price Forecast Models

Finalization of All Models
Offline & Online Testing

Data Sources

Hotel search history*

Hotel id	Search time	Check-in date	Number of nights	Number of adults	Number of children	Base room rice (USD)
123945543	2018-12-03 12:34:45	2019-12-23	2	2	0	358
1024856	2019-03-29 13:34:23	2019-06-23	4	2	1	281
7495038	2019-04-03 10:23:02	2019-05-14	1	1	0	189

*Artificial data example printed for confidentiality purpose

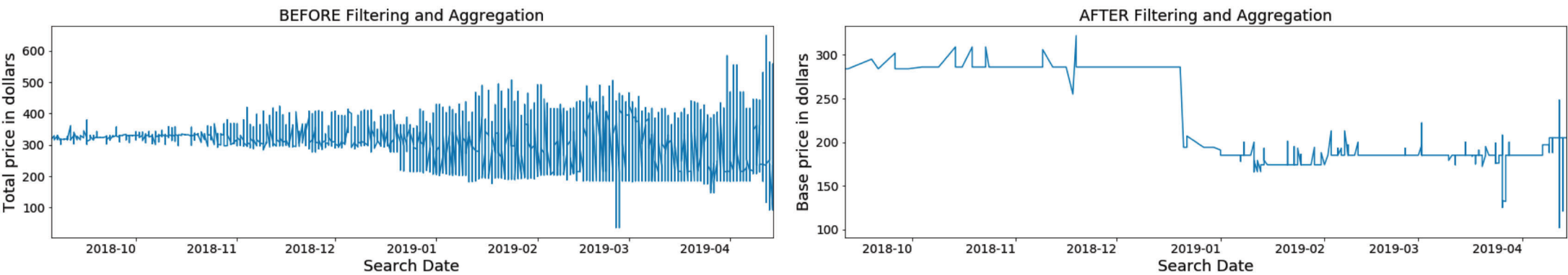
- Scope
- Searches from September 2018 to July 2019

127 external providers

416 hotels (Manhattan)

More than 20B rows

Exploratory Data Analysis



Variation Causes

Solutions: Data transformation

1 External Providers

2 Price Conversion

3 Number of Nights

4 Room Types

Aggregated by hour and day with mode prices

Forecasted on base prices instead of total prices

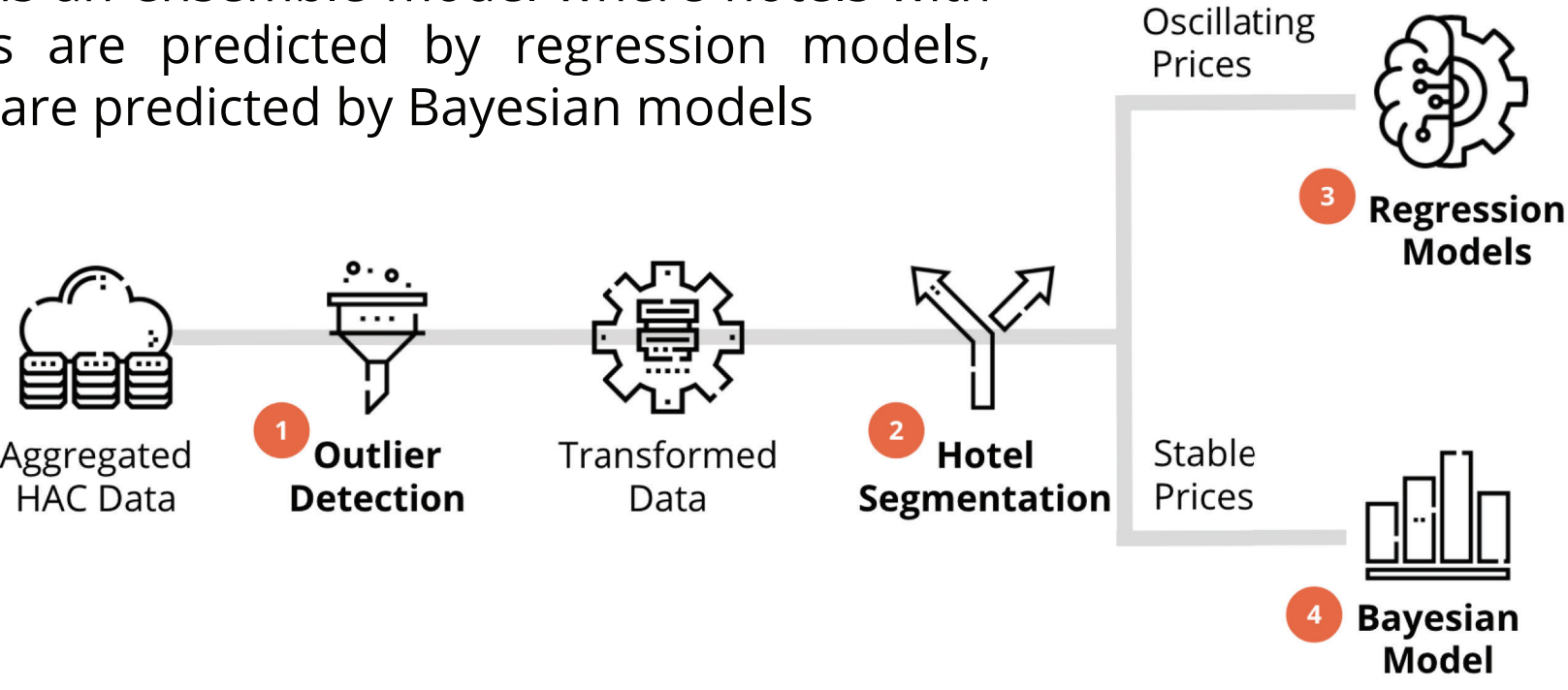
Filtered on USD as original currency

Filtered on number of night equals one

Filtered on queries that have two adults and zero child

Methodology

Our final model is an ensemble model where hotels with oscillating prices are predicted by regression models, while the others are predicted by Bayesian models



1. Outlier detection

Naïve rule

DBSCAN

Gaussian Mixture Models

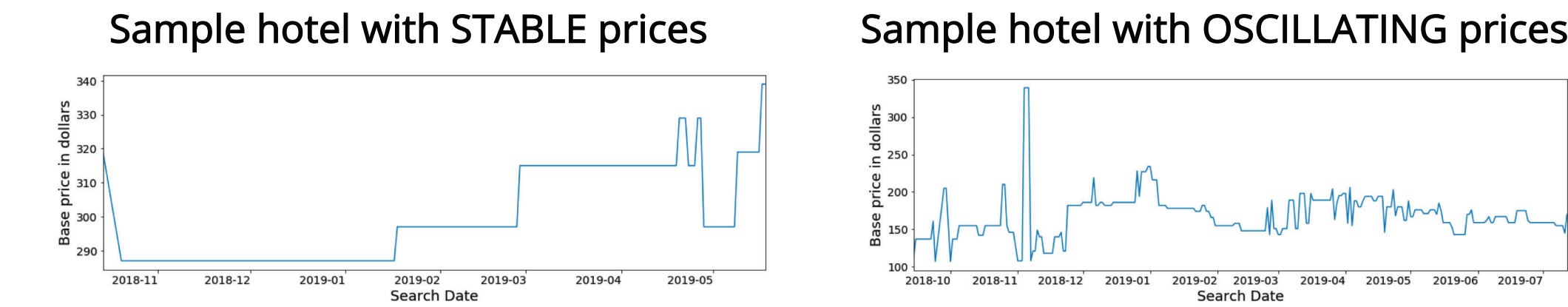
Outliers: Data points greater than median + 4σ of prices on the same check-in date

Outliers: Data points outside density regions

Outliers: Data points that have low likelihood or are in distributions with low soft counts

2. Hotel Segmentation

Powered by K-means clustering using the degree of continuity of price oscillations and the average duration of price plateaus



3. Regression Models

- Features
- Time until travel, decomposed into buckets of time

Day of week of the check-in date

Month of the check-in date

Day of week of the search date

Latest room prices in history

For the requested check-in date and adjacent check-in dates

For the requested check-in date and past search dates

Whether the check-in date is a holiday or not

For the hour aggregated data: hour of the search date

Models

- OLS

Lasso Regression

CART

Ridge Regression

Holistic Regression

4. Bayesian Models

Using Bayes theorem, estimate the rate of price change θ for each hotel, on 2 different time windows

- Long-term - Reserve model: past 30 days prices

Short-term - Vanguard model: past 5 days prices
- $$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

Price emission ~ Bernoulli | Prior and Posterior ~ β | Likelihood ~ Exponential

- 1 Price update: Prior ~ Beta → New price → Update Posterior ~ Beta

2 Hypothesis test: $H_0: P_{reserve}(\theta|X)$ and $P_{vanguard}(\theta|X)$ follow the same distribution

If rejected: replace reserve with vanguard

Otherwise: do not change

3 Prediction: if $P_{reserve}(\theta|X) \geq \tau$, the predicted price is the latest historical price

Final Model Results

For the ensemble model, out-of-sample performances are in average:

	R-squared	MAPE
Baseline*	-0.32	26%
Ensemble model	0.89	3%
Improvement over baseline	+ 1.21**	- 89%***

*Baseline: median hotel price from the past search week
**Improvement in R-square is calculated in relative difference
***Improvement in MAPE is calculated in percent change

The most significant features for regression models are:

Rank	Feature
1	Latest price for the check-in date
2	Latest price from yesterday
3	Latest price from 2 days ago
4	Latest price from 3 days ago
5	Latest price for the day before check-in

Potential Business Opportunities

\$8.2M extra revenue is estimated if our approach is implemented to the following business applications

Hotel sorting

Cheap price alert

Price trend chart

Hotel cross-sell

Heatmap calendar

Next Step



A/B Testing

Conduct an A/B test and assess model and business metrics to prove the approach has positive treatment effects



Implementation

Implement the approach in production



Model Retrain & Monitor

Outlier detection models will update on a monthly basis and forecast models will be monitored and retrain accordingly