Closed Loop Team

Lumen: A Data Science Competition

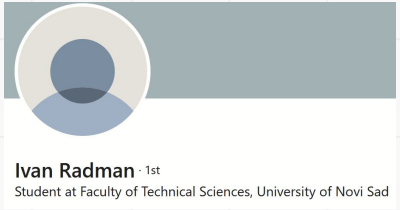
Anja Kovačević
Boris Čuljak
Ivan Radman
Nenad Radović

Our Closed Loop Team

Boris Culjak
Junior Researcher at BioSense Institute

Nenad Radović · 1st
Student at Faculty of Technical Sciences, University of Novi Sad

Anja Kovačević · 1st
Intern at Synchrotek | Student at Faculty of Technical Sciences, Novi Sad



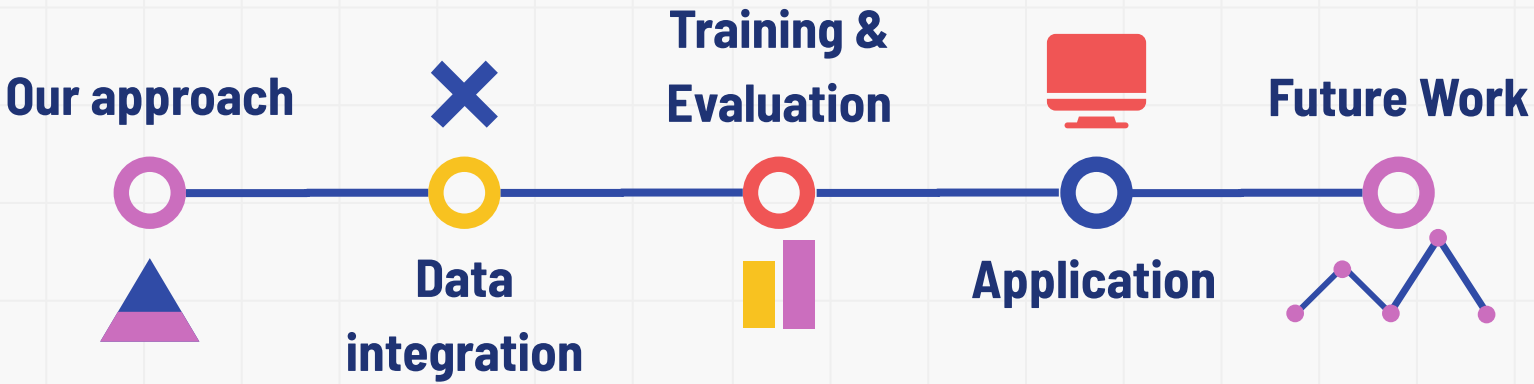
Ivan Radman · 1st
Student at Faculty of Technical Sciences, University of Novi Sad

ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА НОВИ САД

acs
automation & control systems group

- Faculty of Technical Sciences, University of Novi Sad, Serbia
- Control System Engineering

Overview


Our approach
Data integration
Training & Evaluation
Application
Future Work

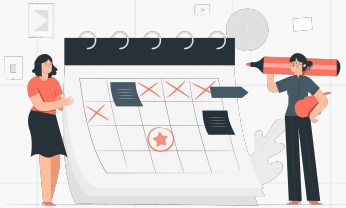
01

Our Approach To The Problem

Task:
In this data science competition, our task was to predict future hotel occupancy rates based on historical occupancy data.

Solution includes:

- Exploration of historical hotel occupancy records
- Constructed a event table capturing significant events and public holidays
- Structured the dataset to align with time series forecasting methods



Initial Dataset

- Alfatec provided the dataset and competition guidelines for this event.
- The analysis focuses on data from the years 2008 and 2009
- Duplicate entries were checked and no missing values were found in critical columns

Column Name	Data Type	Description
reservation_id, guest_id	Integer	Unique IDs for reservation and guest
night_number, room_cnt	Integer	Number of nights and rooms
stay_date, date_from, date_to	datetime64[ns], Object	Dates related to the stay
guest_country_id, reservation_status	Object	Country ID and reservation status
price, price_tax, total_price	Float	Pricing details including taxes
room_category_id, sales_channel_id	Integer, Float	Room category and sales channel

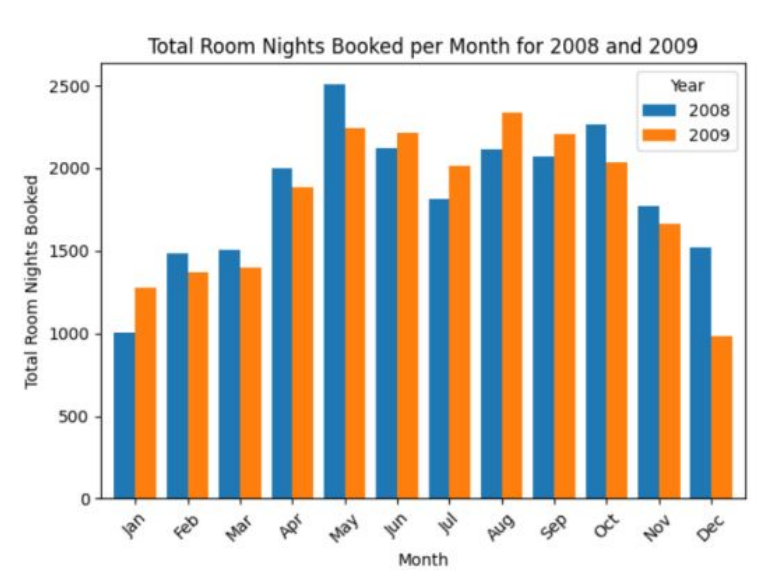
Visits from around the Europe

- The majority of bookings were made by solo travelers or pairs, typically adults.
- Hotel's primary focus is on business conferences, adult leisure, or senior excursions.
- Guests came from 66 different countries, with the top five listed below:

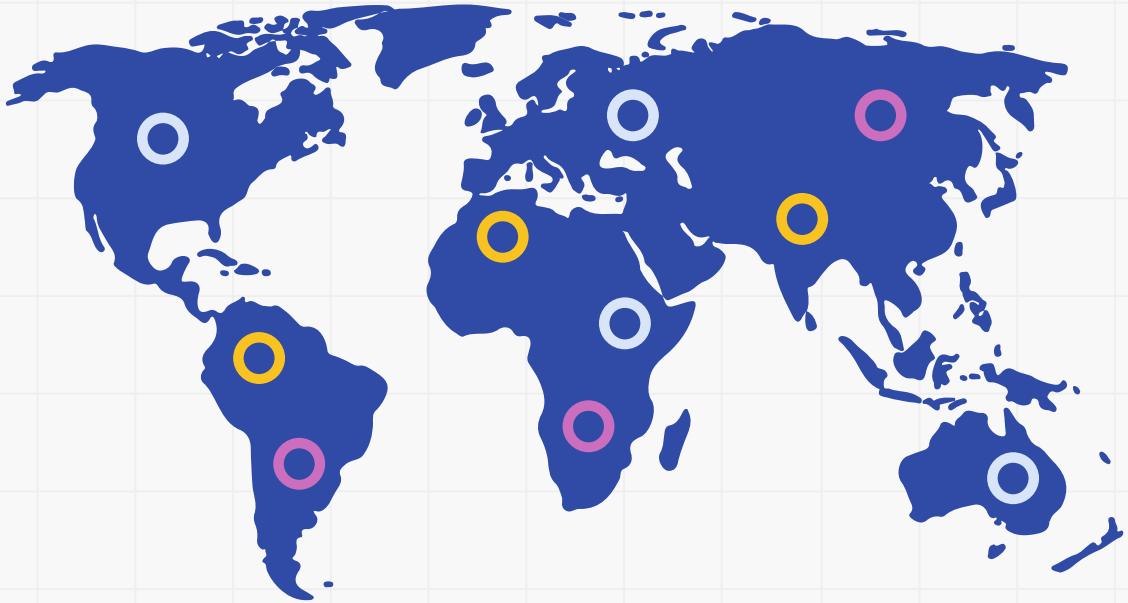
Croatia	Italy	France	UK	Netherlands
16061	6041	2507	1712	1052

Seasonal and Trend Analysis

- Confirmed seasonal demand variations: higher bookings in spring and summer, lower in other months.
- Conclusion: data shows yearly periodicity.



Our addition - Event Table



Pivotal

Impactful component on our problem

Influential

Allows us to analyze patterns

Anticipation

Allowing managers to strategize decisions

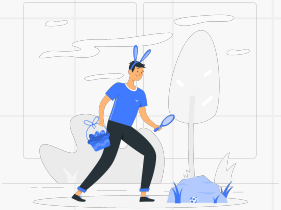
What is in the table?

Event name

The name of the event

Start date

Event start



Type

Possible types of events - festivals, holidays,...

End date

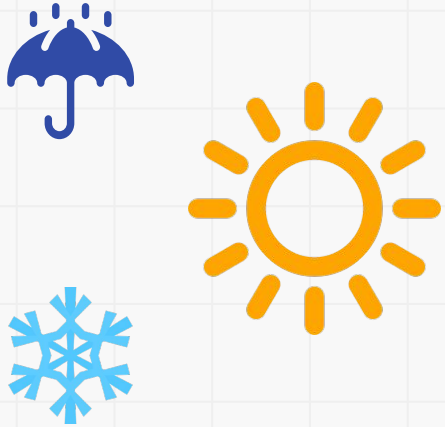
Event finish

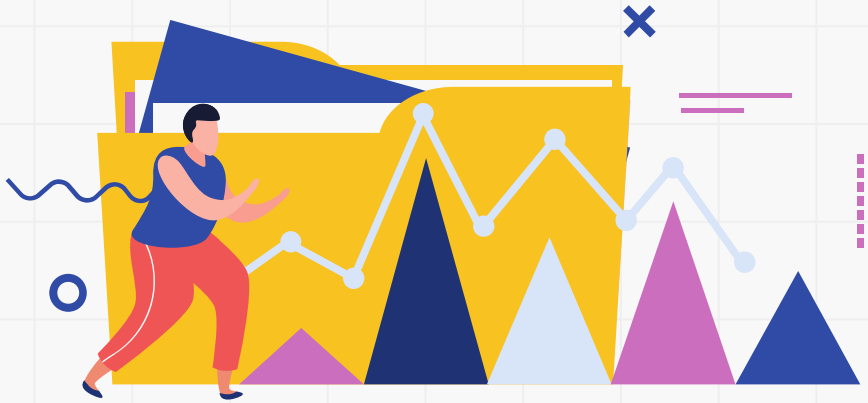
An event example

Event name, type, starting and finishing dates throughout the years...

Weather forecasting data

- Explored the influence of weather on tourist activities and hotel occupancy rates
- Integrated historical weather data from Visual Crossing Weather for Rijeka, Croatia
- Weather data not included in the final dataset but acknowledged for potential relevance





02

Data integration

Data integration

Events

- Filtered by room type due to significant occupancy rate differences
- Retained events where occupancy was 1.5 times higher than the previous month's average

Lagged Data

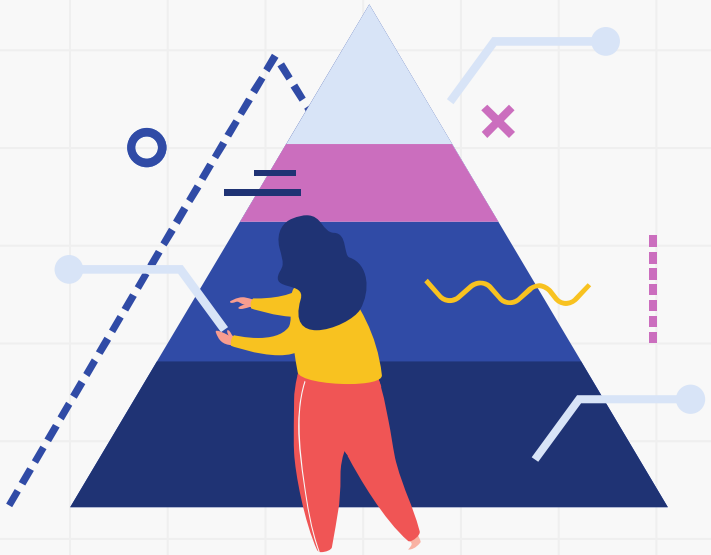
- Utilized occupancy data from the previous seven days
- Each day used as an individual input

Dynamic forecasting model data

Attribute	Description
room_cnt	The number of rooms booked
day_of_week	Day of the week as an integer (e.g., 1 for Monday)
day_of_year	Day of the year, ranging from 1 to 365/366
scaled_room_id	An identifier for each room, scaled between 0 and 1
isEvent	A binary indicator (0 or 1)

Short-term model data

Attribute	Description
day_of_week	Day of the week as an integer (e.g., 1 for Monday)
week_day_avg	Average value of occupancy for that day
month_avg	Average value of occupancy for that month
week_day_importance	Inverse rank of that day importance based on occupancy
event	Indicator for events
occupancy_lag_1	Occupancy of day before
occupancy_lag_2	Occupancy of day before 2 days
occupancy_lag_3	Occupancy of day before 3 days
occupancy_lag_4	Occupancy of day before 4 days
occupancy_lag_5	Occupancy of day before 5 days
occupancy_lag_6	Occupancy of day before 6 days
occupancy_lag_7	Occupancy of day before 7 days
mean_last_7	Mean value of occupancy in last 7 days



03

Experiments, Training and Evaluation methodology

Short-term model features and targets

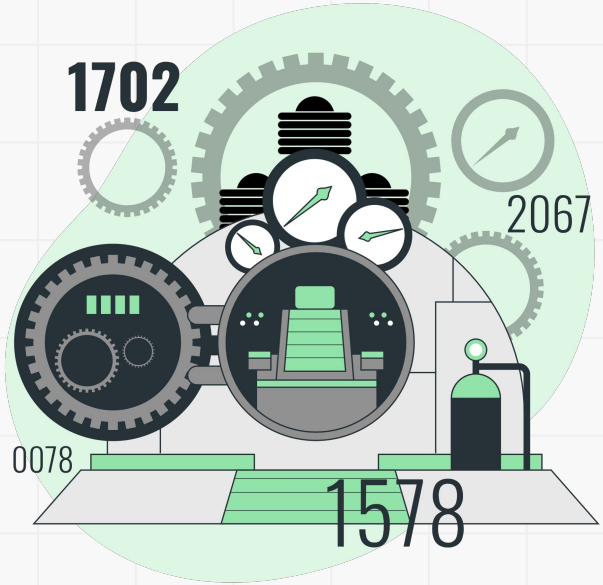
Features selection performed

Types of features:

- Selected events indicators
- Lagged data
- Statistical data

Targets:

- Occupancy for next seven days



Metrics

Why are metrics important?

What do they indicate?

Ones we used:

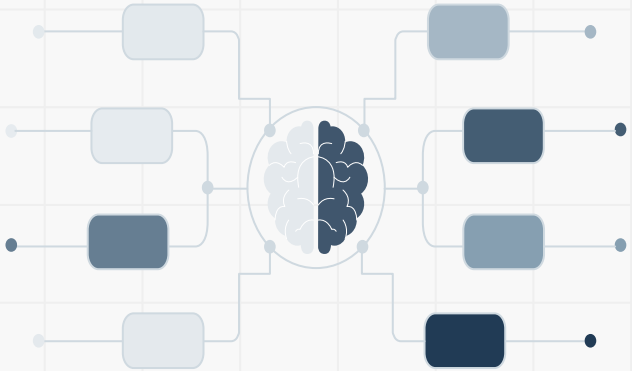
- Mean squared error - MSE
- Mean absolute error - MAE
- R-squared - R2

What was problem and what we concluded?

Model selection

Tested models:

- Linear Regression
- Ridge Regression
- Gradient Boosting
- XGBoost
- Random Forest
- SARIMA



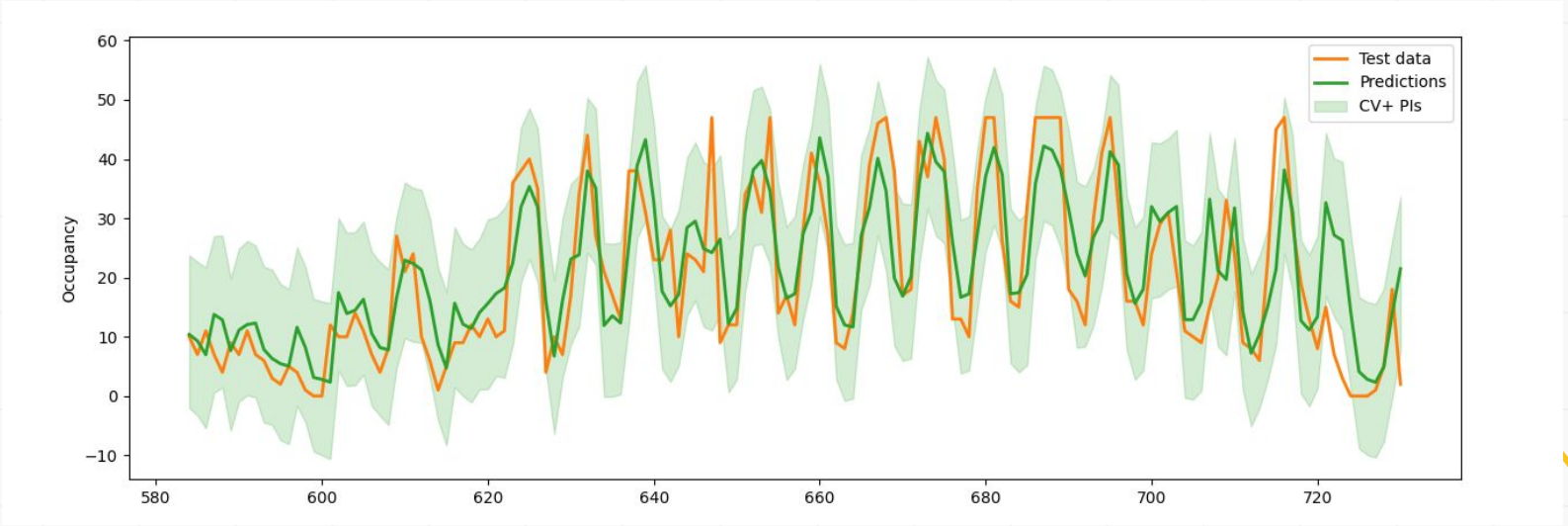
Conclusion:

- Low occupancy rooms » Linear Regression
- High occupancy rooms » Random Forest

Conformal Prediction

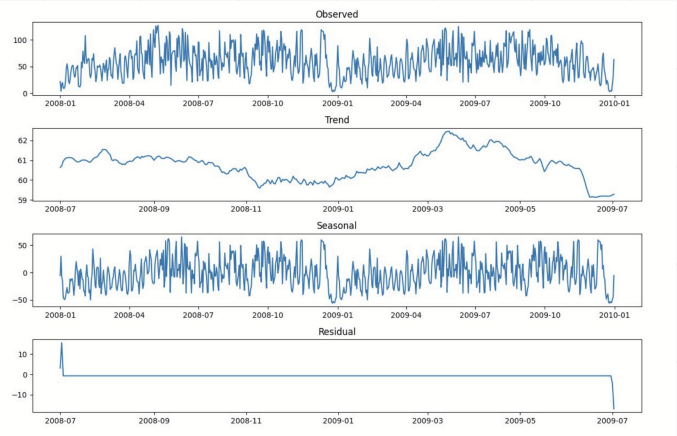
Statistical method of uncertainty quantitation

Why is it important?

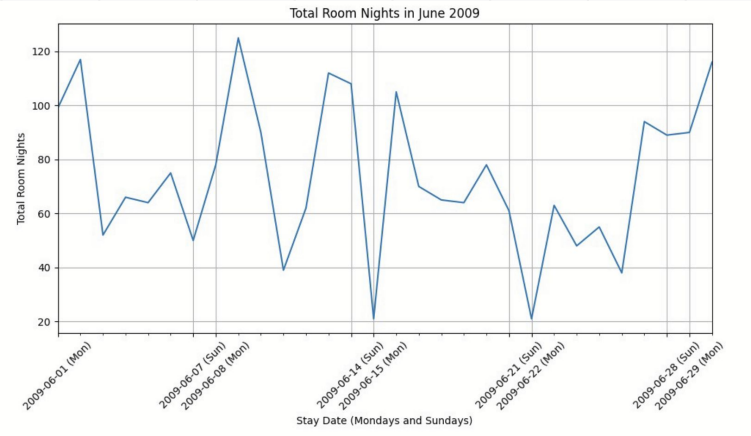


Dynamic forecasting model

- Used for longer time periods
- Leveraged periodic patterns in the data to predict future room bookings.
- Features: Day of the week (1 for Monday to 7 for Sunday), Day of the year (1-365), Room type (scaled category ID), Weather conditions (transformed precip type: 0 for rain, 0.5 for snow, 1 for sun), Event presence (binary indicator for event days)
  - Standardized features to have mean 0 and variance 1.



Full Period



June 2009

day_of_week	day_of_year	scaled_room_id	is_event	weather	room_cnt	eventName
7	6	0.982	1	0	3	Bogojavljanje
7	6	0.337	1	0	1	Bogojavljanje
7	6	1.0	1	0	5	Bogojavljanje
1	7	0.982	1	0	2	Bogojavljanje
1	7	0.337	1	0	2	Bogojavljanje
1	7	1.0	1	0	4	Bogojavljanje
1	7	0.241	1	0	2	Bogojavljanje

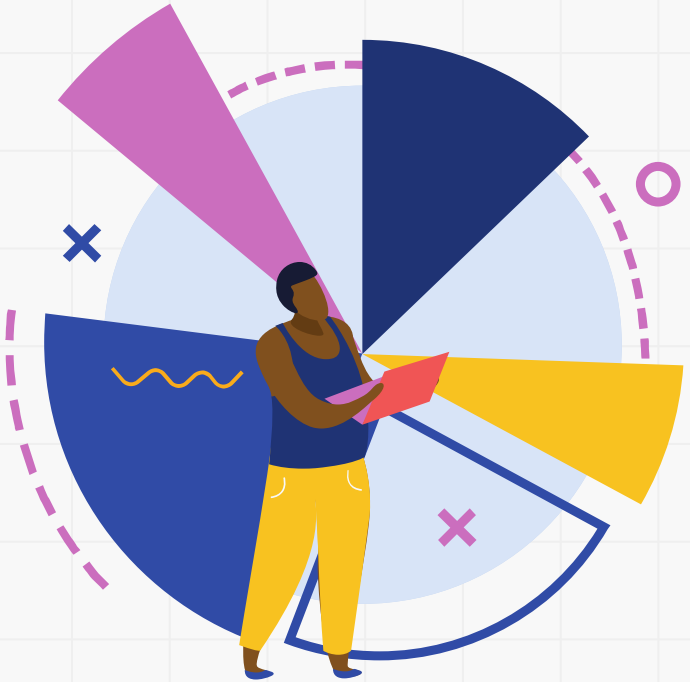
Forecasting

- Train-Test Split: Split data into training (80%) and testing (20%)
- Algorithm: Random Forest Regression
  - Utilized grid search to find the optimal hyperparameter.
  - max_depth=10, min_samples_leaf=2, min_samples_split=5, n_estimators=200.

Metric	Result
MAE	5.683
MSE	89.506
RMSE	9.460
R^2	0.574

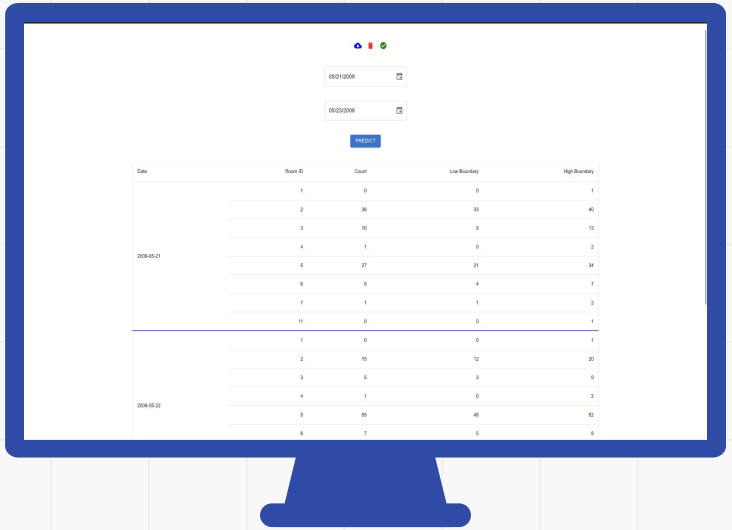
04

Application



Frontend

- React Library
- File Management Section
  - Upload, Delete, Status Indicator
- Date Selection Section
- Prediction Results Table
  - Date, Room ID, Count, Low Boundary, HIgh Boundary



Backend

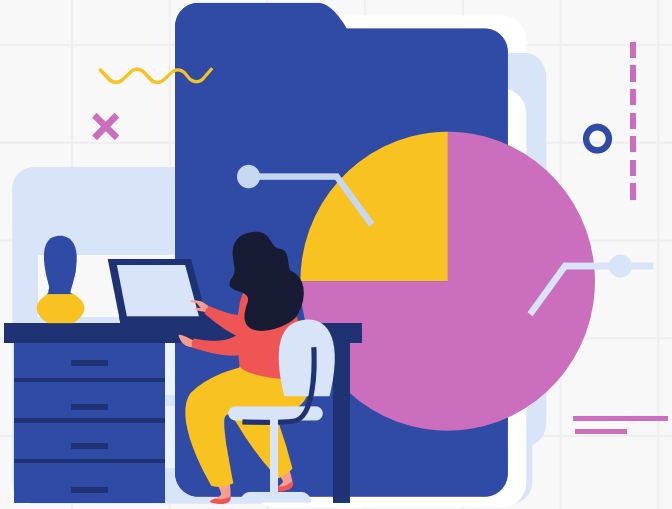
- Flask Framework
- Endpoints
  - CRUD File Operations, Check File Status, Prediction Endpoint
- Processing Functions
  - File validation, Occupancy Calculation, Anomaly Removal, One-Hot Encoding, Feature Engineering, Event Loading, Weather Loading, Normalization…

DevOps

- Complete Dockerization
- Multi-Container Application
  - Frontend Container: Exposed on port 5000
  - Backend Container: Exposed on port 3000
- The application is ready to be deployed
- Build and Run: `docker-compose up --build`

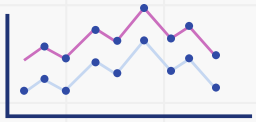
05

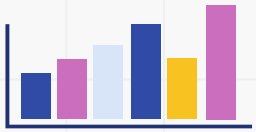
Future Work



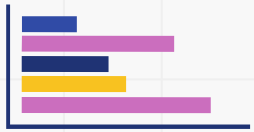
Future work #1

How would we improve our dataset, therefore the whole model





- Enriching dataset
- Integrating Google Trends data
- Conduction of survey
- Travel related data



Future work #2

Improving the overall application experience

1 App UI Design

UI Design features that will amplify hotel management experience

2 Incorporating DL

Advanced Deep Learning methods would help in strategic decision making

3 Database

Storing huge amounts of historical, as well as survey data

4 Hospitality enhancement

Data-driven modern approach for enhancing guests' experiences

Thanks!

Do you have any questions?