|  |  |
| --- | --- |
|  | European Hotel Analysis Presentation Notes |
|  |  |
|  | Jagpreet Bath, Govind Patwal and Ashok (Ash) Parihar |

Overview

[Presenter 1: Jagpreet Bath (Google Slides) 2](#_Toc58163226)

[Project Overview 2](#_Toc58163227)

[Data Source Description 2](#_Toc58163228)

[Questions we are trying answer 3](#_Toc58163229)

[Presenter 2: Govind Patwal (Google Slides) 3](#_Toc58163230)

[Data Preprocessing 3](#_Toc58163231)

[Storing tables in Database 3](#_Toc58163232)

[Presenter 1: Jagpreet Bath (Website) 4](#_Toc58163233)

[Country Dashboard 4](#_Toc58163234)

[Hotels Dashboard 4](#_Toc58163235)

[Review Nationality Dashboard 4](#_Toc58163236)

[Hotels Location Dashboard 4](#_Toc58163237)

[Presenter 3: Ashok (Ash) Parihar (Website) 5](#_Toc58163238)

[Machine Learning Overview 5](#_Toc58163239)

[Logistic Regression Overview 5](#_Toc58163240)

[Ensemble Learners Overview 6](#_Toc58163241)

[Presenter 2: Govind Patwal (Website) 7](#_Toc58163242)

[PySpark NLP 7](#_Toc58163243)

[Deep Learning NLP 8](#_Toc58163244)

[The future of this project. 8](#_Toc58163245)

# Presenter 1: Jagpreet Bath (Google Slides)

## Project Overview

Sally wants to go on vacation. She is thinking about where to go and which hotel should she choose to stay? What are her options?

1. Choose randomly

2. Choose based on experience or recommendation of friends and family

3. Choose based on Hours of Research.

Also, we have faced the same problem. We might have a destination in place but which hotel to stay in is not an easy choice to make. It took hours of research to find a hotel.

The topic we select for our Final Project is a "European Hotel Review Analysis". The objective of our project is to create a web application that can be used by the consumer to find hotels faster and easily, and hotel management can use to improve their performance.

Here are the steps that will allow us to achieve our goal:

1. Create a database and hosted on AWS.
2. Run various machine learning models to predict review scoring and compare which machine learning model is most accurate
3. Create a fully functioning and interactive dashboard using tableau.
4. Create and host a Web application on Github to showcase results.

Lastly, based on previous class deliverables and group research done by the project team, the dataset presents the best opportunity to showcase skills associated with ETL (Extract, Transfer, and Load), machine learning models, and data visualization.

## Data Source Description

Dataset

This dataset contains 515,000 customer reviews and a scoring of 1493 luxury hotels across Europe. The geographical location of hotels is also provided. The data was scraped from Booking.com. All data in the file is publicly available to everyone already. Data is originally owned by Booking.com.

Data Content

The CSV file contains 17 fields. Here is the list of fields.

## Questions we are trying answer

* Train and evaluate various machine learning models to determine which selected machine learning model predicts the review score with the highest accuracy.
* What is the Avg Review Score per Country?
* What % are positive reviews and negative reviews per Country?
* Overtime - Did the Avg Reviews score of the Country improved or degraded?
* What are the Top 5 hotels with the highest review score per Country?
* Overtime - Did the Avg Reviews score of the Top 5 Hotels per Country improved or degraded?
* What is the Number of Reviews per Reviewer Nationality?
* What are the Top 5 Reviewer Nationality per country?
* What % are positive reviews and negative reviews per Top 5 Reviewer Nationality per Country?
* Based on the reviewer's nationality - What are the positive and negative Avg words count per Top 5 Hotels per country?
* What is the recommended hotel per country?

# Presenter 2: Govind Patwal (Google Slides)

## Data Preprocessing

* Step 1: Observe the source data - While observing the Source file: Hotel\_Reviews.csv in a pandas DataFrame
* Data contains 515738 rows × 17 columns
* Step 2: Since there was no unique column in the table, we used the pandas dataframe index to create a unique column 'Review\_Id' for working with data in PostgreSQL
* Step 3: Remove null values
* Step 4: We found that, for best project efficiency, the 17 columns could be divided into two groups: 'review group' and 'location group' with the review group having information pertaining to reviews only, and the location group with information pertaining to location only.
* Step 5: Create sample tables with 1000 rows
* Step 6: Extract city and country using lat and lng, using geopy

## Storing tables in Database

1. PostgreSQL

* Step 1: Create the ERD diagrams using QuickDBD
* Step 2: Create Schemas using QuickDBD
* Step 3: Import the tables into PostgreSQL
* Step 4: View Table 1
* Step 5: View Table 2
* Step 6: Join the 2 tables
* Step 7: View the combined table

2. AWS RDS PostgreSQL database

* Step 1: Creating a PostgreSQL DB on AWS
* Step 2: Connecting the AWS RDS to local PgAdmin

# Presenter 1: Jagpreet Bath (Website)

## Country Dashboard

Country Dashboard allows users to get information related to Countries Hotels. Country Dashboard contains charts with countries hotels Avg Review Score, Avg Positive and Avg Negative word count and % Review Class for each country.

**How will the user use the Dashboard?**

User(s) can select any country from any chart and the Country Dashboard will show information related to the country.

## Hotels Dashboard

Hotels Dashboard allows users to get information related to Hotels in each Country. Hotels by Review Score Dashboard contains charts showcasing Top 5 Hotels by Avg Review Score per country, Top 5 Hotel Positive and Negative Avg word count per country, and Top 5 Hotels by Avg Review Score per country over time.

**How will the user use the Dashboard?**

User(s) can select any country or any hotel from the Top 5 hotels per country from any chart and the Hotels Dashboard will show information related to the country or hotel.

## Review Nationality Dashboard

Reviewer Nationality Dashboard allows users to get information related to the reviewer(s). Reviewer Nationality Dashboard contains charts showcasing Number of Reviews per Reviewer Nationality, Top 5 Review Nationality per country, Top 5 Review Nationality by % Review Class per country, and Review Nationality by Avg Positive Word Count, Avg Negative Word Count per Top 5 Hotels per Country.

**How will the user use the Dashboard?**

User(s) can select any country or any reviewer nationality and the Reviewer Nationality Dashboard will show information related to the country or reviewer nationality.

## Hotels Location Dashboard

Hotels Location Story contains Maps with All Hotel Locations and Top 5 Hotel Location per Country based on the Avg. Review Score.

**How will the user use the Dashboard?**

Allow users to see all the hotel locations and the Top 5 hotel locations per country information.

# Presenter 3: Ashok (Ash) Parihar (Website)

## Machine Learning Overview

Describe Machine Learning: The use of statistical algorithms to perform tasks such as learning from patterns and making predictions.

ML (Machine Learning) Algorithms Used:

1. Logistic Regression (Supervised ML)
2. Random Forest Classifier (Supervised ML)
3. Easy Ensemble Classifier (Supervised ML)
4. Natural Language Processing (NLP)

Describe Supervised ML: Supervised Machine Learning deals with and analyzes labelled data.

Leave Overview webpage and enter Logistic Regression webpage.

## Logistic Regression Overview

Describe Logistic Regression: Algorithm that predicts binary outcomes; there are only two possible outcomes. Common problems in Supervised ML are class imbalances.

One of the objectives of the Machine Learning portion of our project was to predict if hotel reviews provided by customers would either be "positive" or "negative". To address the class imbalance issue, we utilized the following resampling algorithms.

1. Random Oversampling
2. SMOTE
3. Random Undersampling
4. SMOTEENN

Here are the steps that will allow us to code and run the various algorithms: To review these steps, we opened the Repo Link for our Random Oversampling code to display.

* Step 1: Reading our Dataset from our hotel\_combined.csv file
* Step 2: Creating a temporary empty column called Review Class at the end of our Dataset.
* Step 3: Labelling Reviewer\_Scores either Positive or Negative.
* Step 4: Using the LabelEncoder to turn the labels in the Review\_Class column into numbers.
* Step 5: Using the LabelEncoder to turn the labels in the Country column into numbers.
* Step 6: Checking the balance of our target values. There is a significant imbalance in the direction of the positive reviews.
* Step 7: Defining our target variable ('Review\_Class') and excluding unnecessary columns in our Machine Learning Model.
* Step 8: Splitting our data into testing and training sets.
* Step 9: Resampling the training data with our machine learning model of choice (RandomOversampler in this case).
* Step 10: Training the Logistic Regression model using the resampled data.
* Step 11: Calculating the balanced accuracy score.
* Step 12: Displaying the confusion matrix.
* Step 13: Printing the imbalanced classification report.

Note that Step 9 varies based on which resampling algorithm we are running. For the steps listed above, we are coding Random Oversampling.

Other notes from the various steps:

Splitting the data into training and testing sets. The model uses the training dataset to learn from it, then uses the testing dataset to assess its performance.

We randomly oversampled the minority class with the library.

By resampling the dataset, we were able to train the model, make predictions and evaluate the model’s performance.

Summarize Logistic Regression: Simple Logistic Regression, Oversampling, Undersampling, and SMOTEENN are used to address class imbalance scenarios. The balanced accuracy scores for our initial round of logistic regression testing were between 0.52 and 0.95. After adding more data from our dataset for the second round of testing, the balanced accuracy scores ranged from 0.77 to 0.96.

Explain Logistic Regression Recommendations: Including a customer’s average score, as well as their positive and negative word counts provided the machine learning algorithms/models more data in order to make predictions on whether future reviews would be positive or negative. Performing Cluster Centroid Undersampling in addition to Random Undersampling would have provided more comparison prediction data.

Leave Logistic Regression webpage and enter Ensemble Learners webpage.

Ensemble Learners Overview

Explain the differences between Ensemble Learners and Logistic Regression.

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. The two algorithms we used for this project are Random Forest and Easy Ensemble Classifiers.

Just like we did previously with Logistic Regression, we wanted to compare various balanced accuracy scores from different algorithms. Random Forest and Easy Ensemble were two more algorithms for comparison.

Here are the steps that will allow us to code and run the various algorithms: For this explanation, we did not open the Repo Link due to time constraints in addition to having done so once earlier in the presentation.

* Step 1: Reading our Dataset from our hotel\_combined.csv file
* Step 2: Creating a temporary empty column called Review Class at the end of our Dataset.
* Step 3: Labelling Reviewer\_Scores either Positive or Negative.
* Step 4: Using the LabelEncoder to turn the labels in the Review\_Class column into numbers.
* Step 5: Using the LabelEncoder to turn the labels in the Country column into numbers.
* Step 6: Checking the balance of our target values. There is a significant imbalance in the direction of the positive reviews.
* Step 7: Defining our target variable ('Review\_Class') and excluding unnecessary columns in our Machine Learning Model.
* Step 8: Splitting our data into testing and training sets.
* Step 9: Resampling the training data with our machine learning model of choice (BalancedRandomForestClassifier in this case).
* Step 10: Calculating the balanced accuracy score.
* Step 11: Displaying the confusion matrix.
* Step 12: Printing the imbalanced classification report.

Summarize Ensemble Learners: Random Forest and Easy Ensemble Classifier models are also used to address class imbalance scenarios. The balanced accuracy scores for our initial round of logistic regression testing were between 0.59 and 0.65. After adding more data from our dataset for the second round of testing, the balanced accuracy scores ranged from 0.79 to 0.80.

Moving from logistic regression algorithms to Random Forest and Easy Ensemble Classifier algorithms produced higher accuracy scores. This was not a surprise as Random Forest and Easy Ensemble Classifiers combine different subsets of the minority and majority data to generate more accurate predictions.

Explain Ensemble Learners Recommendations: Including a customer’s average score, as well as their positive and negative word counts provided the machine learning algorithms/models more data in order to make predictions on whether future reviews would be positive or negative.

The Easy Ensemble Classifier model produced a slightly higher balanced accuracy Score of 0.795. The Easy Ensemble Classifier algorithm is recommended to use for predicting positive and negative hotel reviews.

Due to time constraints, we are not able to run a K-Means model for a comparison. Comparing supervised with unsupervised machine learning models would have been beneficial.

# Presenter 2: Govind Patwal (Website)

## PySpark NLP

* Step 1: Processing data for Attempt 1
* Step 2: Cleaning data using PySpark in Google Colab
* Step 3: Running NLP (1st Attempt) and observing results
* Step 4: Processing data for Attempt 2
* Step 5: Running NLP (2nd Attempt) and observing results
* Step 6: Processing data for Attempt 3
* Step 7: Running NLP (3rd Attempt) and observing results
* Step 8: Processing data for Attempt 4
* Step 9: Running NLP (4th Attempt) and observing results
* Step 10: Processing data for Attempt 5

Summary of Results

The accuracy of the model has been

1. For output as decimals (0.0 to 10.0): 20.77 %
2. For output as integers (0 to 10): 39.62%
3. For output as 5-star ratings (1 to 5): 61.27%
4. For output as 3 Categories (Bad, Average, or Good): 77.63%
5. For output as Sentiment (Positive or Negative): 91.56%

## Deep Learning NLP

1. Using the input as reviews and output as decimal score (0.0 - 10.0)
2. Using the input as reviews and output as integers (0 - 10)
3. Using the input as reviews and output as 5 stars (1 - 5)
4. Using the input as reviews and output as 3 Categories (Good, Average, Bad)
5. Using the input as reviews and output as Sentiments (Positive/Negative)

Summary of Results

1. For output as decimals (0.0 to 10.0): 0.0 %
2. For output as integers (0 to 10): 0.0 %
3. For output as 5-star ratings (1 to 5): 0.43 %
4. For output as 3 Categories (Bad, Average, or Good): 3.10%
5. For output as Sentiment (Positive or Negative): 95.66 %

Recommendation(s)

* The DL NLP is a very accurate model if a user wants to predict reviewer sentiments.
* It is, however, a big failure for predicting reviews in any other of the other 4 formats - the PySpark NLP fares much better.

## The future of this project.

POA

1. Capture every review.
2. Use the saved Deep Learning NLP model to predict the sentiment of the review.
3. Add the data (review + sentiment) to our PostgreSQL Database on RDS
4. Use Tableau to analyze the data
5. The data upload and analysis can be scheduled once a week/month, or it can happen real-time on the cloud.

POC (review capture and sentiment prediction)

1. Open Google Colab
2. Sign in with your Google credentials
3. Upload this file on Google Colab
4. Place this file at the root of the Colab file above
5. Run the File: Runtime -> Run All
6. Enter as many reviews as you want to test
7. View the model predictions.

Sample Outcome