**PROJECT REPORT**

(Project Term August-December 2021)

## Prediction of Hotel Booking-Cancellation

Submitted by

**Pratyush Pranjal Registration Number: 11906033**

**Nishant Registration Number: 11910329**

**Course Code INT246**

Under the Guidance of

**Dr. Sagar Pande**

**Assistant Professor**

# School of Computer Science and Engineering,

# Lovely Professional University, Punjab

A picture containing text

Description automatically generated

**DECLARATION:**

We hereby declare that the project work entitled **Prediction of Hotel Booking-Cancellation** is an authentic record of our own work carried out as requirements of project for the award of B. Tech. degree in CSE from Lovely Professional University, Phagwara, under the guidance of Dr. Sagar Pande , during August to December 2021. All the information furnished in this project report is based on our own intensive work and is genuine.

Name of Student 1: **Pratyush Pranjal**

Registration Number: **11906033**

Name of Student 2: **Nishant**

Registration Number: **11910329**

**CERTIFICATE:**

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Project under my guidance and supervision. The present work is the result of their original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech. degree in CSE from Lovely Professional University, Phagwara.

**Dr. Sagar Pande**

**Designation: Assistant Professor**

**School of Computer Science and Engineering,**

Lovely Professional University,

Phagwara, Punjab.

Date : 20 November 2021

**ACKNOWLEDGEMENT:**

This project is the result of our project group and would not have been completed properly without the combined efforts put in by its members. So, we are thankful to each other for the proper co-operation and industrious work hours put in by each of us.

We are highly indebted to Dr. Sagar Pande for his guidance and constant supervision, helping us out with each complication that we ran into as well as for providing necessary information regarding the project along with the tedious support in completing the project.

We would like to express our gratitude towards our parents, our family and friends for their kind co-operation and encouragement regardless of their degree of assistance which helped us in completion of this project.

At last, we would like to thank our university for providing us the opportunity to opt for such a project as our in-course project and gain all the knowledge and experience that we could gather.

Name of Student 1: Pratyush Pranjal

Registration Number: 11906033

Name of Student 2: Nishant

Registration Number: 11910329

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **I** | **TITLE PAGE** | **(i)** |
| **II** | **DECLARATION** | **(ii)** |
| **III** | **CERTIFICATE** | **(iii)** |
| **IV** | **ACKNOWLEDGEMENT** | **(iv)** |
| **V** | **TABLE OF CONTENT** | **(v)** |
| **1.** | **INTRODUCTION** | **7** |
| **2.** | **PROBLEM STATEMENT** | **8** |
| **3.** | **EXISTING SYSTEM ANALYSIS**  **3.1 INTRODUCTION**  **3.2 EXISTING SYSTEM**  **3.3 DATA FLOW DIAGRAM**  **3.4 WHAT’S NEW** | **9 - 10** |
| **4.** | **DATASET DESCRIPTION**  **4.1 DATA DICTIONARY** | **11 - 12** |
| **5.** | **SYSTEM REQUIREMENT ANALYSIS**  **5.1 TOOLS**  **5.1.1 SOFTWARE REQUIREMENT**  **5.1.2 HARDWARE REQUIREMENT**  **5.2 BUDGET**  **5.3 SPECIFIC REQUIREMENT**  **5.3.1 JUPYTER NOTEBOOK**  **5.3.1.1 NUMPY**  **5.3.1.2 PANDAS**  **5.3.1.3 MATPLOTLIB**  **5.3.1.4 SCIKIT LEARN**  **5.3.2 GOOGLE COLAB** | **13 - 18** |
| **6.** | **6. MACHINE LEARNING ALGORITHMS**  **6.1 LOGISTIC REGRESSION**  **6.2 NAÏVE BAYES**  **6.3 RANDOM FOREST**  **6.4 DECISION TREE**  **6.5 KNN**  **6.6 SVM** | **19 - 25** |
| **7.** | **IMPLEMENTATION** | **26** |
| **8.** | **PROJECT LEAGACY**  **8.1 CURRENT STATUS OF THE PROJECT**  **8.2 REMAINING AREA OF CONCERN**  **8.3 TECHNICAL AND MANAGERAIL LESSON LEARNT** | **27** |
| **9.** | **SOURCE CODE** | **28 – 32** |
| **10.** | **RESULT ANALYSIS** | **33** |
| **11.** | **APPLICATION OF THE PROJECT** | **34** |
| **12.** | **FUTURE WORK** | **35** |
| **13.** | **CONCLUSION** | **36** |
| **14.** | **BIBLIOGRAPHY** | **37** |
| **15.** | **REFERENCE** | **38 - 39** |

**1. INTRODUCTION:**

We belong to a century where everything is expected to be as efficient as possible. It is known that everyone, irrespective of their field of work, wants to maximize their profits and avoid the factors that can jeopardize the expected result. Here, in this project, we focus upon a field of specific services which is Hotel Bookings. Hotel Bookings is a vast field to focus upon. There are numerous factors that affect the profits from this area of services that affect the yield at a large as well as small scale. As of now, we focus upon the most strongly affecting factor, Hotel Booking-Cancellation.

One of the biggest problems and challenges facing the hospitality industry is the significant number of canceled reservations. Common reasons for cancellations include a sudden deterioration in health, accidents, bad weather conditions, schedule conflicts, or unexpected responsibilities. Interestingly, a noticeable group consists of customers who, after making a reservation, are still looking for new, better offers, and even make many reservations at the same time to be able to choose the most advantageous one.

Being provided with a dataset about the concerning topic enables us to use the power of machine learning at its best. Here, we can use multiple machine learning algorithms and compare which algorithm provides us the best accuracy according to our dataset.

With the use of machine learning, we not only identify the potential factors that lead to a booking cancellation but also predict whether a particular booking is going to cancelled or not.

There are different machine learning algorithms which can be used for the purpose of solving this problem statement. We will mainly focus on algorithms like Logistic Regression, KNN, Random Forest, Decision Tree and Naïve Bayes. These are just few of the algorithms which we are going to implement and with the help of these, we are going to derive as much as trends as we can and further use them to predict the respective outcomes with the efficiency that we achieve.

**2. PROBLEM STATEMENT:**

Over the years, the hotel industry has encountered several problems that affect their profits up to a great scale. If researched extensively, we could shed some light on factors that affect their revenue quite extensively. One of the major factors that cause such an adverse effect is Booking-Cancellations.

Booking Cancellations are one of the primary causes of revenue damage that the hotel industry face nowadays. There are numerous reasons which cause a booking cancellation like health issue, weather, climate, geographical location of the hotel, the seasons, etc. It is understandable that cancellations would ultimately result in financial damages, which is the primary consequence but what people fail to understand is that it causes an effect on different aspects as well. In addition to the direct financial consequences of cancellations, they also cause operational problems (such as over or understaffing). Those problems may lead to decrease customer satisfaction and negative reviews. In a world where more and more customers check online reviews before picking a hotel, those reviews can have major impacts. At a single hotel level, an increase in online reputation score has been linked to an increase in occupancy and revenue. We can clearly understand why avoiding bad reviews due to a room not being ready when the guest arrives can be very valuable for a business. This requires knowing which booking to prioritize. It is therefore very useful for hotels to know which bookings are likely to get canceled in order to plan their operations accordingly.

In order to solve this problem, we will use a real-life hotel booking dataset to create a customer segmentation analysis in order to gain insights about the customers (and hopefully reasons why they cancel their reservation). We will then build a classification model (including the newly created customer clusters) to predict whether or not a booking will be canceled with the highest accuracy possible.

This model will allow hotels to predict if a new booking will be canceled or not, manage their business accordingly, and increase their revenue.

**3. EXISTING SYSTEM:**

3.1 INTRODUCTION:

Hotel Booking system is complex, multi-level system that involves multiple external entities, numerous processes, plenty of data stores and a tremendous amount of data flow. Customers normally go through different processes and confirm whether they want to have the booking done or they prefer to cancel and modify as per their requirement.

3.2 EXISTING SYSTEM:

Like most of the systems, this one also starts with the customer when they enquire about room availability. When done with this phase, they are asked to purchase the booking which generates booking details, received by both hotels and customers too after the next step is done successfully, which is the payment. All the while this information and statistics are being submitted to different data stores for proper functionality. And at last, we have the customer reviews for proper selection next time.

3.3 DATA FLOW DIAGRAM:

Confirmation Details
Enquiry Details
Customer
Room availability Details
Room
availability
Check Room
Details
1
Room
Purchase

3.4 WHAT’S NEW:

This projects now comes right after the booking confirmation. When the customers confirm their bookings by completing the payment process, this project would intervene. With this project in the picture, we thereby begin to start analysis whether a particular booking done by a customer would ultimately result in cancellation or not.

As far as booking systems are concerned, we know that after cancellation, it takes a while just to update the scenario on their database. This, at the end, would harm the hotel company as the other customers would see the rooms occupied, although the actual situation is not so. After the implementation of this model, we would increase the efficiency of the system up to a great extent. After a booking the is done, this model will compare all the stats and try and predict whether it would cancel or not before it gets cancelled. This would ultimately benefit all parts of the hotel industry.

**4. DATASET DESCRIPTION:**

It contains information about bookings from two hotels in Portugal for the period from July 2015 to August 2017. One of the hotels is a resort hotel and the other is a city hotel. There are 119 390 observations in total (each of them describes one reservation). The key is that about 37% of them were cancelled which is a pretty large number. The dataset provides us multiple features to use which are as following:

4.1 DATA DICTIONARY:

|  |  |
| --- | --- |
| is\_canceled | Value indicating if the booking was cancelled or not |
| adr | Average Daily Rate, calculated by dividing the sum of all lodging transactions by the total number of staying nights |
| adults | Number of Adults |
| agent | ID of the travel agency that made the booking |
| arrival\_date\_week\_number | Week number of the arrival date |
| arrival\_date\_year | Year of arrival date |
| arrival\_date\_month | Month of arrival date |
| booking\_changes | Number of changes/amendments made to the booking from the moment the booking was entered on the Property Management System (PMS) until the moment of check-in or cancellation |
| country | Country of Origin |
| contract | when the booking has an allotment or other type of contract associated to it |
| deposit\_type | Indicating if a guarantee deposit was made the booking. |
| children | Number of children |
| stays\_in\_weekend\_nights | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| stays\_in\_week\_nights | Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| meal | Type of meal booked |
| market\_segment | Market segment designation |
| distribution\_channel | Booking distribution channel |
| is\_repeated\_guest | Indicating if the booking name is from a repeated guest |
| previous\_cancellations | Number of previous bookings that were cancelled by the customer prior to the current booking |
| previous\_bookings\_not\_canceled | Number of previous bookings not cancelled by the customer prior to the current booking |
| reserved\_room\_type | Code of room type reserved. Code is presented instead of designation for anonymity reasons. |
| booking\_changes | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation |
| lead\_time | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| required\_car\_parking\_spaces | Number of car parking spaces required by the customer |
| total\_of\_special\_requests | Number of special requests made by the customer |
| reservation\_status | Reservation last status |
| reservation\_status\_date | Date at which the last status was set |

**5. SYSTEM REQUIREMENT ANALYSIS:**

5.1 TOOLS:

For application development, the following Software Requirements are: Operating System: Windows 7 or any Linux Debian Distro.  
Language: R and Shiny  
Tools: RStudio IDE, Microsoft Excel (Optional).

Technologies used: R, Unix, Shiny.

Operating System Network  
Visio Studio

GitHub  
Google Chrome

5.1.1 SOFTWARE REQUIREMENTS:

Any OS with clients to access the internet Wi-Fi Internet or cellular Network

Create and design Data Flow and Context Diagram  
Versioning Control  
Medium to find reference to do system testing, display and run shiny App.

5.1.2 HARDWARE REQUIREMENTS:

For application development, the following Software Requirements are: Processor: Intel or high  
RAM: 1024 MB  
Space on disk: minimum 100mb

For running the application:  
Device: Any device that can access the internet Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORETM i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

5.2 BUDGET:

The budget of completion for developing the hotel booking cancellation prediction system will require various software and hardware devices. The application is averagely expensive to build but if happens to be as successful as the developer sees it to be it will bring forth enough profit to cover the costs undergone.

5.3 SPECIFIC REQUIREMENTS:

5.3.1 JUPYTER NOTEBOOK:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebook documents. The “notebook” term can colloquially refer to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context.

According to the official website of Jupyter, Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

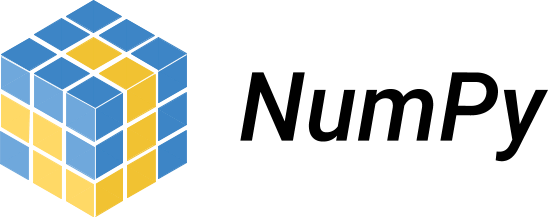
Jupyter Book is an open-source project for building books and documents from computational material. It allows the user to construct the content in a mixture of Markdown, an extended version of Markdown called MyST, Maths & Equations using MathJax, Jupyter Notebooks, reStructuredText, the output of running Jupyter Notebooks at build time. Multiple output formats can be produced (currently single files, multipage HTML web pages and PDF files).

5.3.2 GOOGLE COLAB:

Just like Jupyter Notebook, google colab is an online substitute of Jupyter where everything can be done just Jupyter.

5.3.1.1 NUMPY:

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

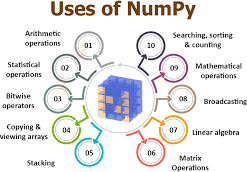


At the core of the NumPy package, is the ndarray object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance. There are several important differences between NumPy arrays and the standard Python sequences:

* NumPy arrays have a fixed size at creation, unlike Python lists (which can grow dynamically). Changing the size of

an ndarray will create a new array and delete the original.

* The elements in a NumPy array are all required to be of the same data type, and thus will be the same size in memory. The exception: one can have arrays of (Python, including NumPy) objects, thereby allowing for arrays of different sized elements.
* NumPy arrays facilitate advanced mathematical and other types of operations on large numbers of data. Typically, such operations are executed more efficiently and with less code than is possible using Python’s built-in sequences.
* A growing plethora of scientific and mathematical Python- based packages are using NumPy arrays; though these typically support Python-sequence input, they convert such input to NumPy arrays prior to processing, and they often output NumPy arrays. In other words, in order to efficiently use much (perhaps even most) of today’s scientific/mathematical Python- based software, just knowing how to use Python’s built-in sequence types is insufficient - one also needs to know how to use NumPy arrays.



5.3.2 PANDAS:

The Pandas library is one of the most important and popular tools for Python data scientists and analysts, as it is the backbone of many data projects. Pandas is an open-source Python package for data cleaning and data manipulation. It provides extended, flexible data structures to hold different types of labelled and relational data. On top of that, it is actually quite easy to install and use.

Pandas is often used in conjunction with other data science Python libraries. In fact, Pandas is built on the NumPy package, so a lot of the structure between them is similar. Pandas is also used in SciPy for statistical analysis or with Matplotlib for plotting functions. Pandas can be used on its own with a text editor or with Jupyter Notebooks, the ideal environment for more complex data modelling. Pandas is available for most versions of Python, including Python3.



Think of Pandas as the home for your data where you can clean, analyze, and transform your data, all in one place. Pandas is essentially a more powerful replacement for Excel. Using Pandas, you can do things like:

* Easily calculate statistics about data such as finding the average, distribution, and median of columns
* Use data visualization tools, such as Matplotlib, to easily create plot bars, histograms, and more
* Clean your data by filtering columns by particular criteria or easily removing values
* Manipulate your data flexibly using operations like merging, joining, reshaping, and more
* Read, write, and store your clean data as a database, txt file, or CSV file

5.3.3 MATPLOTLIB:

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object- oriented API for embedding plots into applications using general- purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.[3] SciPy makes use of Matplotlib.

There are numerous plots that can be performed using matplotlib. During visualization of the dataset, matplotlib acts as a great medium to help users identify pattern and analyze datasets as much as possible. There are some instances where different plots play a more significant role than as compared to the other ones.

5.3.4 Scikit-Learn:

Scikit-learn is largely written in Python and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible.

Scikit-learn integrates well with many other Python libraries, such as Matplotlib and plotly for plotting, NumPy for array vectorization, Pandas dataframes, SciPy, and many more.

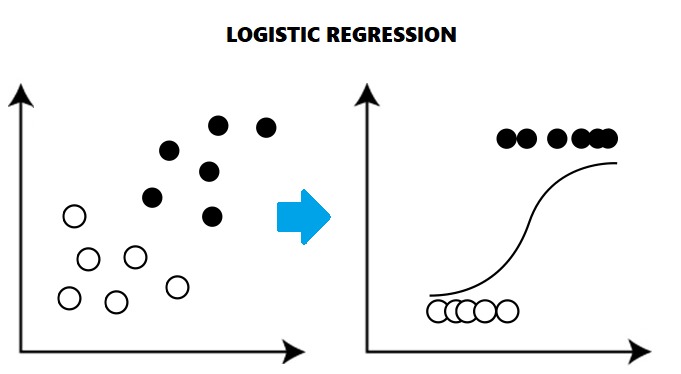
**6. MACHINE LEARNING ALGORITHMS:**

6.1 Logistic Regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.



## Types of Logistic Regression:

Binary or Binomial

In such a kind of classification, a dependent variable will have only two possible types either 1 and 0

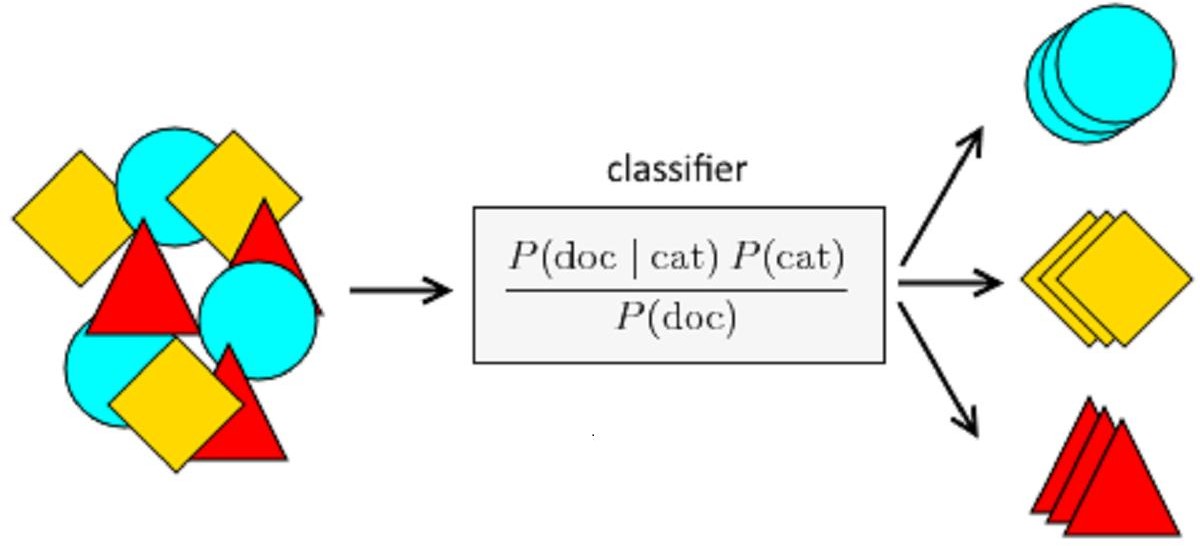
Multinomial:

In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance.

Ordinal

In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance.

6.2 Naïve Bayes:

Naïve Bayes algorithms is a classification technique based on applying Bayes’ theorem with a strong assumption that all the predictors are independent to each other. In simple words, the assumption is that the presence of a feature in a class is independent to the presence of any other feature in the same class. For example, a phone may be considered as smart if it is having touch screen, internet facility, good camera etc. Though all these features are dependent on each other, they contribute independently to the probability of that the phone is a smart phone.

In Bayesian classification, the main interest is to find the posterior probabilities i.e. the probability of a label given some observed features, P (L | features). With the help of Bayes theorem, we can express this in quantitative form as follows −

P(L|features)=P(L)P(features|L)P(features)P(L|features)

=P(L)P(features|L)𝑃(𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠)

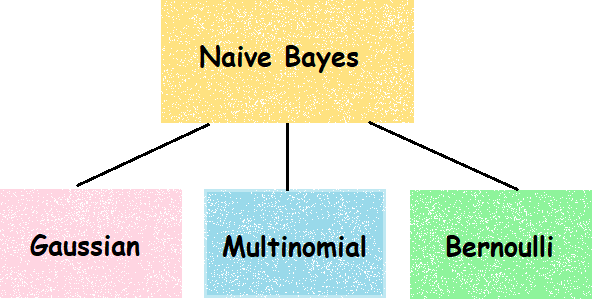
Here, 𝑃(𝐿 | 𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠) is the posterior probability of class.

𝑃(𝐿) is the prior probability of class.

𝑃(𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠 | 𝐿) is the likelihood which is the probability of predictor given class.

𝑃(𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠) is the prior probability of predictor.

## TYPES OF NAÏVE BAYES:



Gaussian Naïve Bayes

It is the simplest Naïve Bayes classifier having the assumption that the data from each label is drawn from a simple Gaussian distribution.

Multinomial Naïve Bayes

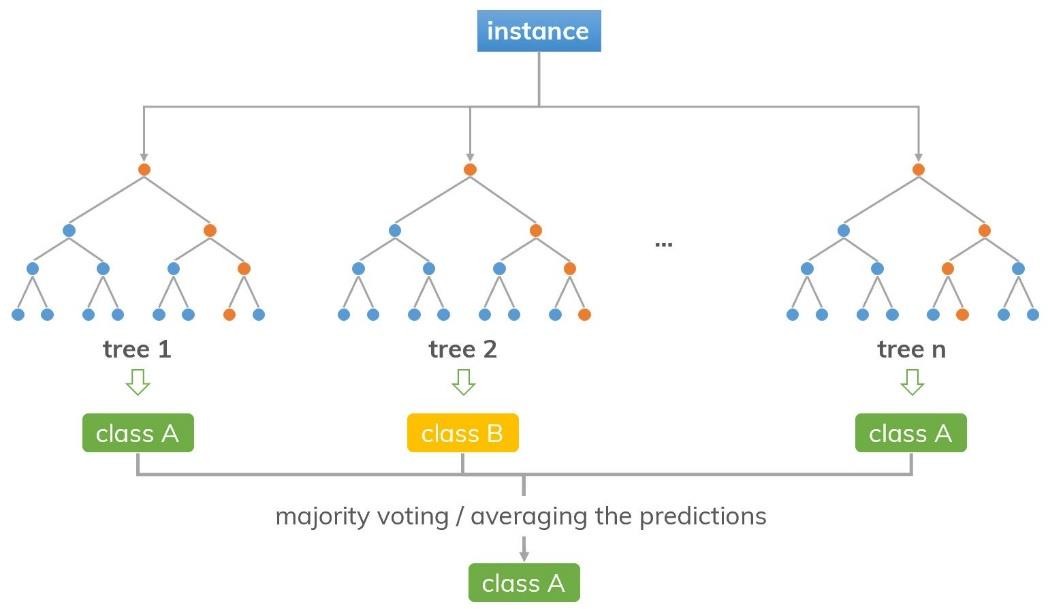
Another useful Naïve Bayes classifier is Multinomial Naïve Bayes in which the features are assumed to be drawn from a simple Multinomial distribution. Such kind of Naïve Bayes are most appropriate for the features that represents discrete counts.

Bernoulli Naïve Bayes

Another important model is Bernoulli Naïve Bayes in which features are assumed to be binary (0s and 1s). Text classification with ‘bag of words’ model can be an application of Bernoulli Naïve Bayes.

6.3 Random forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Applications of Random Forest:

There are mainly four sectors where Random Forest mostly used:

1. Banking: Banking sector mostly uses this algorithm for the identification of loan risk.

2. Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.

3. Land Use: We can identify the areas of similar land use by this algorithm.

4. Marketing: Marketing trends can be identified using this algorithm.

6.4 Decision Tree:

A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules.

Diagram

Description automatically generated

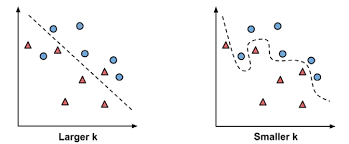
Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are non-parametric supervised machine learning method used for both **classification**and **regression**tasks. the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Figure 9.1 describes a decision tree that reasons whether or not a potential customer will respond to a direct mailing. Internal nodes are represented as circles, whereas leaves are denoted as triangles. Note that this decision tree incorporates both nominal and numeric attributes. Given this classifier, the analyst can predict the response of a potential customer (by sorting it down the tree) and understand the behavioral characteristics of the entire potential customers population regarding direct mailing. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding value

6.4 kNN:

Machine learning techniques have been widely used in many scientific fields, but its use in medical literature is limited partly because of technical difficulties. k-nearest neighbors (KNN) is a simple method of machine learning. Here, we introduce some basic ideas underlying the KNN algorithm, and then focuses on how to perform KNN modeling with python. The dataset should be prepared before running the KNN function. After prediction of outcome with KNN algorithm, the diagnostic performance of the model should be checked. Average accuracy is the mostly widely used statistic to reflect the KNN algorithm. Factors such as k value, distance calculation and choice of appropriate predictors all have significant impact on the model performance.

The impact of selecting a smaller or larger K value on the model

* Larger K value**:** The case of underfitting occurs when the value of k is increased. In this case, the model would be unable to correctly learn on the training data.
* Smaller k value**:**The condition of overfitting occurs when the value of k is smaller. The model will capture all of the training data, including noise. The model will perform poorly for the test data in this scenario.



Here are some things to keep in mind:

1. As we decrease the value of K to 1, our predictions become less stable. Just think for a minute, imagine K=1 and we have a query point surrounded by several reds and one green (I’m thinking about the top left corner of the colored plot above), but the green is the single nearest neighbor. Reasonably, we would think the query point is most likely red, but because K=1, KNN incorrectly predicts that the query point is green.
2. Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (up to a certain point). Eventually, we begin to witness an increasing number of errors. It is at this point we know we have pushed the value of K too far.
3. In cases where we are taking a majority vote (e.g. picking the mode in a classification problem) among labels, we usually make K an odd number to have a tiebreaker.

6.4 Support Vector Machine:

Support Vector Machine(SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.



**7. IMPLEMENTATION:**

This project was implemented with the use of Jupyter Notebook, where we imported all the necessary libraries like numpy, pandas, matplotlib, seaborn, etc. The first and foremost step was to read the CSV(comma separated value) file. After doing that, we moved on to the next phase of building a machine learning model, which is Exploratory Data Analysis. While performing EDA, we follow the trends and examine the dataset to answer questions to go through the data thoroughly. In this phase, we answer questions like which month produces a great number of customers and that too in which hotel, which hotel receives the majority number of guests, where are these customers from, which country has the majority of the cancellations and derive patterns.

These patterns are further utilized to make predictions, for e.g. looking at the trends, it can easily be derived that majority of the cancelled bookings belong to the same country where the hotels in our dataset are situated in. The international bookings rarely get cancelled as compared to domestic bookings. This trend is just an example of how a particular prediction could be figured out.

If we look at the overall methodology of machine learning, the primary objective is to gather the data which we already have in the form of CSV file. After this, comes the data preparation, as mentioned above, where we prepare the data to perform specific tasks and functions and derive some new trends and patterns. This is now followed by removing inaccurate data from the dataset and then transforming the same data, typically, by changing “raw” data into another format which is much more suitable for use. After this, we analyse the dataset and deduce all the trends and pattern. After this, we build a model and train it with either a part of dataset or we train it with an entire dataset and later on, test it with a different one. If we get a good accuracy, we move ahead and deploy the model.

**8. PROJECT LEGACY:**

8.1 CURRENT STATUS OF THE PROJECT:

This project is, as of now, ready for the state of deployment. We have our model up and running. According to the dataset we had, we have used multiple machine learning algorithms to check which one provides us the best accuracy. We concluded that the ML algorithm called KNN gave us the best accuracy of 94.8% and the rest had mixed accuracies. Some were close enough to compete with KNN, but few were far from comparison as well.

8.2 REMAINING AREAS OF CONCERN:

We know that this project only concerns with predicting information on the basis of the dataset that we have at our end, which is actually dated to 2015 to 2017. There are a lot of factors that have not yet been included in this dataset. So, real-time dataset would definitely help the hotel industry or the organization using this model. In addition, we can also say that this model has been made for machine learning enthusiasts. We are yet far from creating this for proper public use. Therefore, with the use of different software like HTML, CSS and JavaScript or Flask or Django, we can create a web UI for the same and avail the services publicly.

8.3 TECHNICAL AND MANAGERIAL LESSONS LEARNT:

After performing this project, one thing became certain that machine learning models can be beneficial up to a great extent, irrespective of the field or industry it belongs to. With the use of ML algorithms, we intend to help out the industry in ways that can avoid extra revenue loss and also help in proper managerial matters, which could be caused due to constant booking cancellations. In conclusion, we would say that if we can get ours hands on a better, more proficient our model would come out to be.

**9. SOURCE CODE:**

Importing necessary libraries and reading the data:

**Table

Description automatically generated**

Data Information: Data Cleaning:

**Graphical user interface, table

Description automatically generated Table

Description automatically generated**

Treating Redundant values:

**Table

Description automatically generated**

EDA Patterns and Trends:

**Graphical user interface, map

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart

Description automatically generated with medium confidence**

**Chart

Description automatically generated**

**10. RESULT ANALYSIS:**

COMPARE ALL MODEL

Table

Description automatically generated**Table

Description automatically generated**

**A screenshot of a computer

Description automatically generated with low confidenceTable

Description automatically generated**

**Table

Description automatically generated** **Table

Description automatically generated**

**As we can see that the dataset that have is best suited for KNN, Random Forest or Decision Tree, particularly in that order.**

**11. APPLICATIONS OF THE PROJECT:**

This project can be applicated in numerous fields and industries. We know that the creation of this model was for the sole purpose of improving and expediting the hotel industry by helping them patch the loophole of showing bookings unavailable as there are customers waiting for a chance to book a room but due to lack in efficiency or unforeseen cancellations, they are not able to do so. But with the help of this model, we can increase proficiency of the system and help out a lot of clients by avoiding loss of revenue.

There are other fields also where this type of ML model would be of great assistance, some of them are listed as following:

1. Restaurant Booking System:

Since, this field of industry is not that different from Hotel Booking system. The same model can easily expedite the proper efficiency of the system. In addition, the busy restaurants usually face shortage of table bookings, where this model could come in quite handy.

2. Travel Industry:

If we are provided with a proper dataset that has important features with good corelations, we can easily predict booking cancellations and make it available for proper public use.

3. Health Industry:

If we look at situations where patients make bookings, i.e., appointments, we can use this model to predict whether the customer would likely show up to the clinic or hospital.

**12. FUTURE WORK:**

This project is just the beginning of the topic. There are many other fields where this project can be further enhanced and used to perform multiple other tasks. For e.g., we can easily create a standalone application to further expedite the prediction process, where in the application would easily detect which customer are most likely to drop their reservations. By pursuing this approach, there will be no human interference at all. The booking would be easily checked on its own and the user would directly get notified whether the customer is likely to cancel the reservation or not.

With a little change, this can further be used in different areas like the restaurant system where they detect a booked table would get cancelled or not. Same thing goes banquet halls. This can easily be used to predict whether a particular event booking would be performed or not. All in all, there are numerous such cases where this would be quite convenient.

**13. CONCLUSION:**

After going through this project, we can clearly see that machine learning models can be easily created with the use of multiple machine learning algorithms. With the use of all the machine learning algorithms, we can even check which algorithm suits our dataset the best. We check their accuracies; we can perform further changes by putting in extra effort in data wrangling and cleaning to gain a much better accuracy and deploy the model so as to make sure that the users get to use this to the best of its capability.

After performing everything, we concluded that multiple algorithms gave us quite good accuracies like KNN, Random Forest and Decision Tree and there were other algorithms as well, which gave comparatively, a lot less accuracies like Naïve Bayes, Logistic Regression.

After successful implementation, we can easily figure out ways to publicize this model to make sure that common people can use this model to the best of its capability.

**14. Bibliography**

1. <https://chegg.com>

2. [https://ai.google](https://ai.google/)

3. [https://towardsdatascience.com](https://towardsdatascience.com/)

4. [https://youtube.com](https://youtube.com/)

5. [https://www.tensorflow.org](https://www.tensorflow.org/)

6. [https://www.javatpoint.com](https://www.javatpoint.com/)

7. [https://www.stackoverflow.com](https://www.stackoverflow.com/)

8. [https://www.quora.com](https://www.quora.com/)

9. [https://www.analyticsvidhya.com](https://www.analyticsvidhya.com/)

**15. REFERENCE:**

DFD Picture:

<https://media.cheggcdn.com/media/2ab/2ab0a235-61d7-4a13-aae9-925514212e2c/phppjwtRE.png>

Jupyter Notebooks:

<https://towardsdatascience.com/everything-you-need-to-know-about-jupyter-notebooks-10770719952b#:~:text=The%20Jupyter%20Notebook%20is%20an,machine%20learning%2C%20and%20much%20more>.

NumPy Pictures:

[https://miro.medium.com/max/1050/1\*cyXCE-JcBelTyrK-58w6\_Q.png](https://miro.medium.com/max/1050/1*cyXCE-JcBelTyrK-58w6_Q.png)

<https://i0.wp.com/techvidvan.com/tutorials/wp-content/uploads/sites/2/2020/07/Uses-of-NumPy-1.jpg?w=828&ssl=1>

Pandas Picture:

<https://cdn.analyticsvidhya.com/wp-content/uploads/2018/03/pandas-768x480.jpg>

Linear Regression Information:

<https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/>

Logistic Regression Information:

<https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148#:~:text=Logistic%20Regression%20is%20a%20Machine,on%20the%20concept%20of%20probability.&text=The%20hypothesis%20of%20logistic%20regression,function%20between%200%20and%201%20>.

Logistic Regression Picture:

[https://miro.medium.com/max/903/1\*BrLF7ySJ\_6Lfw2IXmNjsXQ.png](https://miro.medium.com/max/903/1*BrLF7ySJ_6Lfw2IXmNjsXQ.png)

Naïve Bayes Information:

<https://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4>

Naïve Bayes Pictures:

<https://insightimi.files.wordpress.com/2020/04/unnamed-1.png>

[https://miro.medium.com/max/1754/1\*-ILVYuyxT3ypZ9CF0BEgWQ.png](https://miro.medium.com/max/1754/1*-ILVYuyxT3ypZ9CF0BEgWQ.png)

Random Forest Information and Picture:

<https://www.javatpoint.com/machine-learning-random-forest-algorithm>

Decision Tree Information and Picture:

<https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96>

kNN Information:

<https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

SVM Information and Pictures:

https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/#:~:text=%E2%80%9CSupport%20Vector%20Machine%E2%80%9D%20(SVM,both%20classification%20or%20regression%20challenges.&text=Support%20Vectors%20are%20simply%20the,hyper%2Dplane%2F%20line).