Loan approval, bike sharing and hotel reviews analysis using different machine learning methods

Valentina Bernal Gomez
Student number: 23395745
Email: x23395745student.ncirl.ie

I. SECTION A: LOAN APPROVAL PREDICTION

***Abstract*—Loan approval prediction is a classification problem in which we need to decide whether the loan will be approved or not. Lenders must always know how a borrower intends to spend a loan and make sure it is not for a purpose they do not allow. Furthermore, they will not approve a personal loan for an illegal activity or risky purpose, the purpose may also influence your personal loan interest rates; typically the lender is a bank that decides whether to grant the loan or not using different aspects, this aspects assesses the ability and willingness of a borrower to fully repay a loan on time.**

For this analysis, four different methods were used to compare and find the best measures, Random Forest method achieved the best accuracy with 92.96% in comparison with Logistic Regression, K-Nearest Neighbors (KNN) and Support Vector Machine.

Keywords—loan approval, classification, machine learning, Random Forest.

A. INTRODUCTION

The loan approval process is a critical step in a financial institution lending decisions to customer. Loans are commonly offered by banks, credit unions, and other financial institutions, and can be used for major purchases, investment, renovations, debt consolidation and business ventures. They provide a flexible solution for those in need of financing. However, loans also come with important considerations, including interest rates, repayment terms, and potential consequences for defaulting on payments.

Loans are typically categorized as either secured or unsecured and can have fixed or variable interest rates, depending on the terms of the loan agreement. A borrower's creditworthiness points to whether a lender can trust the borrower will repay the loan within the specified terms without defaulting.

There are several factors that lenders consider when deciding if a particular borrower is worth the risk, such as income, credit score and debt-to-income-ratio. For larger loans, lenders may require a certain income threshold, by ensuring that the borrower will have no trouble making payments. They may also require several years of stable employment, especially in the case of home mortgages. A credit score is a numerical representation of a person's creditworthiness, based on their history of borrowing and repayment. Missed payments and bankruptcies can cause serious damage to a person's credit score [1]. In addition to one's income, lenders also review the borrower's credit history to check how many active loans they have at the same time. A high level of debt indicates that the borrower may have difficulty repaying their debts [2].

Currently, there are loan software that is used to automatized and processed the approval based on the rules that are set by the lender. However, when defining these rules or criteria it is essential to understand what are the socioeconomic and financial factors that significantly influence the likelihood of loan approval, on the other hand, we will discuss through this report how effective is the supervised model Random Forest in loan approval and which factors are critical when approved a loan, those aspects will be developed through this report.

B. RELATED WORK

The loan approval decision is critical for both banks and borrowers. Traditionally this decision was based on basic statisical models or manual evaluation. but currently different models of machine learning are implemented to know which one is the most suitable for this decision. However, with the increase of available data and machine learning advances there are more automatized and improved models which help whether a loan is approved or not.

Different research has been done by implementing various machine learning models. Random forest is the model with the best accuracy in different reports. Sinap [3] demonstrated a significantly improved model performance using random forest methodology, there he found that models built with selected features achieved higher accuracy, recall, precision and F1 score values than models build with K-nearest neighbors or with all features. The Random Forest algorithm showed the highest performance with an accuracy rate of 97.71%. This study restates the outcome obtained in this report, where Random Forest was the model with high accuracy, feature selection was performed during his report as well in this report, where person age and length of the applicant's credit history were dropped due to high correlation between variables. In terms of which factors are relevant, Vahid Sinap found that married individuals, high income individuals, males and university graduates are more likely to be approved for loans, this finding cannot be supported with our report as we do not have variables such as married individuals or gender.

Deborah et al. [4] utilized three different algorithms such as K-Nearest Neighbors, Decision Tress and Support vector Machine (SVM) to predict loan approval, they concluded the best model with the highest accuracy (83%) for their values was Support Vector Machine, this study differs from ours as they do not use Random Forest, in our case the accuracy reach with SVM is 88%, this could be due to the different implemented factors. They also found features as gender, marital status, number of dependents, self-employment status, applicant income, loan amount, credit history and property location influence their SVM model.

Uddin et al. [5] applied more than five machine learning methods, some of them were not included in our report, such as Extra Trees, AdaBoost and advanced deep learning (deep neural networks, recurrent neural networks and long short-term memory), moreover, they include an ensemble voting model, which increase the accuracy compared to the Extra Tress achieving 87.25% for this measure. Uddin et. al. performs an extra model that we do not include in our report.

A comparative study using decision trees and random forest had been completed by Madaan et al. [6] where they found Random Forest outperformed the Decision Tree with an accuracy of 80%, this confirm and support our study where Random Forest reach the best accuracy.

Dansan et al. [7] analyzed the impact of different loan features on bank loan prediction, this research used different variables than the ones used for this report, such as gender, educational qualification and marital status, the used Random Forest method found that marital status is an essential factor when approving a loan, however, in our study we cannot evaluate this impact as the variable is not included.

C. Data Mining Methodology

To answer the research question "how effective is the supervised model Random Forest in loan approval and which factors are critical when approved a loan?" I used KDD (Knowledge Discovery in Databases), this methodology is a systematic process that seeks to identify useful data, this method was apply as follows:

Data selection: this dataset was selected from Kaggle and contains 12 columns and 32580 rows.

Data pre-processing: The columns 'person employment length' and 'loan interest rate' contains missing values, considering that, those values have been filled using the median, due to is an appropriate measure to fill in missing values when dealing with skewed distributions or when outliers are present in the data.
The correlation matrix was analyzed as shown below:

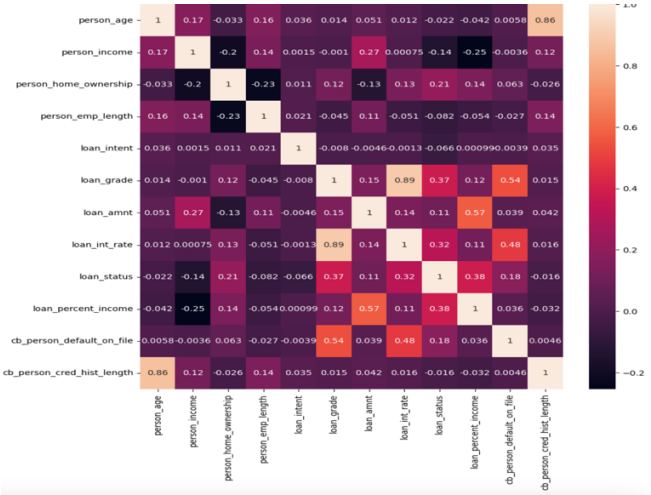


Figure 1. Correlation matrix

From the table above, correlation can be identified, in this case person's age and length of the applicants credit history are correlated with a value of 0.86, for this reason, the column 'person's age' was dropped, considering that the length of the applicants credit history is more important in the loan approval as give key information to the bank about how a person managed their financial obligations during a period of time, for instance a longer and positive credit history is more reliable for the lenders.
Loan interest rate and loan grade appeared to be highly correlated with a value of 0.89, so loan interest rate was removed, given that loan grade represents the borrower's risk indicator.

Data transformation: Normalizing some variables was an important step to ensure the features were in a comparable scale and high variables will not domain other during the modeling, this is useful to preserve the shape of the original distribution, these variables were: income, employment length and loan amount.
Through this step, encoding was developed in the categorical variables, such as, person home ownership, loan intention, loan grade and person default on file, this was done to enable machine learning models works with variables as all the variables has to be numerical.

Data mining: Four machine learning methods (KNN, Logistic Regression, Random Forest and SVM) were analyzed for this dataset, the idea is to compare different measures, as the accuracy, precision, recall and F1-score, to select the best and effective method.

The dataset was split into training (80%) and test (20%) sets to assess how effectively our machine learning model works.

After building the models, in the case of KNN, 'Grid Search' was used to find the hyperparameters, to assess the best value for the number of neighbors.

Data evaluation: This step is the last one used in KDD methodology, and it will be discussed in detail in section D.

D. Evaluation

To answer the research question, "How effective is the supervised model Random Forest in loan approval, and which factors are critical when approving a loan?", the machine learning models mentioned before were trained. Each model was evaluated using performance measures to ensure effectiveness and reliability. Accuracy measures overall performance, Precision measures positive prediction quality, Recall evaluates sensitivity to positive cases and F-1 score shows the balance between Precision and Recall. Using the KDD methodology, I focused on data mining and data evaluation to determine how effective is every model to predict loan approval.

1. Logistic regression: was trained, and the summary model is shown below:

```
Optimization terminated successfully.
         Current function value: 0.367231
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:                26064
Model:                          Logit   Df Residuals:                    26054
Method:                           MLE   Df Model:                            9
Date:                Sun, 17 Nov 2024   Pseudo R-squ.:                  0.2984
Time:                        10:05:18   Log-Likelihood:                -9571.5
converged:                       True   LL-Null:                       -13643.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
=======
                                coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-------
const                         -4.2446      0.075    -56.468      0.000      -4.392
-4.097
person_income                  2.8606      1.912      1.496      0.135      -0.887
6.608
person_home_ownership          0.2735      0.015     18.812      0.000       0.245
0.302
person_emp_length             -0.7441      0.607     -1.226      0.220      -1.934
0.446
loan_intent                   -0.1374      0.011    -12.879      0.000      -0.158
-0.117
loan_grade                     0.8820      0.019     47.228      0.000       0.845
0.919
loan_amnt                     -2.8884      0.144    -20.120      0.000      -3.170
-2.607
loan_percent_income           11.6954      0.240     48.634      0.000      11.224
12.167
cb_person_default_on_file     -0.1087      0.048     -2.279      0.023      -0.202
-0.015
cb_person_cred_hist_length     0.0019      0.005      0.414      0.679      -0.007
0.011
==============================================================================
```

Figure 2. Logistic Regression model

The following measures were done for the model:

```
Logistic Regression:
Accuracy: 0.8375019180604573
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.95      0.90      5072
           1       0.72      0.44      0.55      1445

    accuracy                           0.84      6517
   macro avg       0.79      0.70      0.72      6517
weighted avg       0.83      0.84      0.82      6517

Confusion Matrix:
 [[4821  251]
 [ 808  637]]
```

Figure 3. Results for the Logistic Regression model

The accuracy for the model is 84% for the predictions, however we must analyze the results for the class 0 and 1:

Class 0 (approval): Precision: predictions for the loan approval were 86% correct. Recall: the model detects 95% of all classes that are from class 0

Class 1: Precision: just the 72% were correct for this class, it might have high false positives (i.e. predict class 1 when were class 0). Recall: the model just detects 44% of real cases for this class, it might be a lot of false negatives (i.e. predicts class 0 when were class 1). F-1 score: 0.55, the model is having problems to classify this class.

The model is right for the class 0 but I having issues predicting class 1.

As some p-values shown not significant for some variables, I will drop columns such as person income, employment length and credit history length, to see the impact on the model. The outcome after removing these columns is shown below:

```
Optimization terminated successfully.
         Current function value: 0.367292
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:                26064
Model:                          Logit   Df Residuals:                    26057
Method:                           MLE   Df Model:                            6
Date:                Sun, 17 Nov 2024   Pseudo R-squ.:                  0.2983
Time:                        10:05:19   Log-Likelihood:                -9573.1
converged:                       True   LL-Null:                       -13643.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
=======
                                coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-------
const                         -4.2353      0.064    -65.666      0.000      -4.362
-4.109
person_home_ownership          0.2756      0.014     19.216      0.000       0.248
0.304
loan_intent                   -0.1374      0.011    -12.883      0.000      -0.158
-0.116
loan_grade                     0.8823      0.019     47.261      0.000       0.846
0.919
loan_amnt                     -2.8295      0.132    -21.368      0.000      -3.089
-2.570
loan_percent_income           11.5871      0.224     51.733      0.000      11.148
12.026
cb_person_default_on_file     -0.1092      0.048     -2.289      0.022      -0.203
-0.016
==============================================================================
```

Figure 4. Logistic Regression model

The correspond measures for the model:

```
Logistic Regression:
Accuracy: 0.8368881387141323
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.95      0.90      5072
           1       0.72      0.44      0.54      1445

    accuracy                           0.84      6517
   macro avg       0.79      0.69      0.72      6517
weighted avg       0.82      0.84      0.82      6517

Confusion Matrix:
 [[4819  253]
 [ 810  635]]
```

Figure 5. Results for the Logistic Regression model

As we notice from the results, the measures are the same, the final accuracy for logistic regression after removing some variables is 84%.

2. KNN:

Accuracy: 87%, the model correctly classifies 87% of the total cases. Recall: the model detects correctly 55% of the positive real cases. F1-score: the model is good classifying correctly positive cases but is not excellent. When doing Grid Search the best neighbors is 5, the model considers the 5 nearest neighbors to classify a new point.

3. Random Forest:

This model reaches an accuracy of 92.96% and precision of 96.18%, this indicates that the model correctly classifies approval and rejections. F1- score: 81.69% this shows a good balance between precision and recall. The model handles False Positives, this is essential to do not approve incorrectly a loan.

4. SVM:

This model reaches an accuracy of 88.78%, meaning that the model correctly classifies the 88% of the total cases. F1-score: the value of 75.06% suggests a good performance. Recall: The model detects 76.12% of the True Positive cases (approvals).

As we confirm from the literature revised, Random Forest performs well in this case of loan approval as performed in other reports, showing a high effectiveness to predict loan approval. As we selected Random Forest, we evaluated feature importance, to see which variables are relevant for our model.

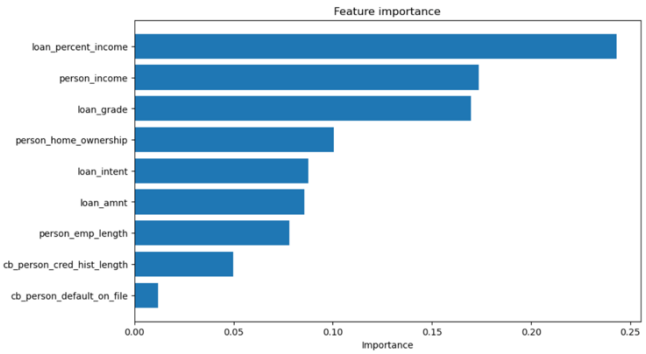


Figure 6. Feature importance

This help us to answer which factors are relevant when approving a loan. Loan percent income and person income are the variables with the highest importance for the model.

E. Conclusions and Future Work

After modelling the different machine learning methods, the best model with the best measures is Random Forest (Accuracy: 92.96%, F1-score: 0.81 and Recall: 0.71), the results suggests that Random Forest is effective at handling complex relationships between the variables and making predictions more accurate. Moreover, it ability to prioritize the approval and rejection cases makes it ideal for loan approval systems. While KNN and SVM also performs well, their efficiency is lower compared to Random Forest.

Despite the good results, there are some limitations that are important to consider, the variables used in the model are relevant, however, we might include more socioeconomic features, such as marital status, gender and number of dependents, that could improve the model.
Although we split the data into training and test sets, the loan approval and rejection data could be unbalanced, and this might impact in the prediction.

For future work, it would be interesting to use deep learning machine techniques and compare their results with the current models. Additionally, build the model using new features will increase the accuracy and measures for the model.

II. SECTION B: BIKE SHARING PREDICTION

Abstract— Bike sharing services have become popular in different cities around the world, this is because it helps people to optimize the time making journey from point A to point B shorter, moreover, contribute to traffic solutions, more people using these services make decrease the use of transport public or the cars. However, forecasting the demand for these services is a challenge considering different factors such as weather, seasons, temperature, etc. This report will develop a regression model to predict the rented bikes per hour.

The findings revealed that Random Forest achieved the best result and is the most effective model, with the lower MSE, MAE and the highest R2.

Keywords—bike sharing, Random Forest, regression model.

A. Introduction

For many big cities around the world, where traffic is an issue and the rush hours have become a main problem due to the spending time in buses or cars, bike sharing schemes have become a solution for this issue. This service is a short-term bicycle rental, where you can pick up the bike at any station and returned to any other station, this works because the service is connected to an app, and you can find the nearest station to your location and then returned in the closest station you would need.

The first generation of Bike Share Schemes (BSS) began in Amsterdam in 1965. Known as “white bicycles” they were to be used for a single trip, then left unlocked for the next person to use. This BSS in Amsterdam helped to increase the number of cyclists, especially among those that did not have a bicycle [8].

Those schemes were introduced in many other countries and were adapted to their necessities and implemented some improvements, for instance Copenhagen gave the possibility to lock the bike and use a system of “coins refund” in order to provide a better service, moreover, in Rennes, France they implemented a smart card access, automatic docks and stations, real time information and “first 30 minutes for free” [8].

Bike sharing systems are reducing congestion in various European cities. Those systems are open to the public for an affordable cost, it is possible to pay for an annual membership, or even for one day pass, allowing tourists to explore cities at a low cost. The usage of bike sharing systems depends on many factors, among them weather plays an important role when people decide to use a bike from this scheme, in theory, bad weather results in less bike usage, by understanding how people rent a bike depending on weather conditions this report could provide useful insights to know how do weather conditions and hourly usage patterns influence demand for bicycle rental systems in urban areas.

B. Related Work

Currently, with the rapid increased of the cities, traffic is becoming a significant problem and the impact to the environment is growing rapidly. Carbon emissions due to the use of cars are turning into a concern. Bike sharing services are helping to improve commuting time and provide an option to use an alternative mode of transportation which do not exacerbate this environment problem. The use of this service could have a higher demand in cities where the weather is warmer or clearer. There are many studies who suggests the use of machine learning models and there are some who use predictive models more advanced such as neural networks, Pan et al. [9] proposes a model based on long short-term memory to predict the rents and returns of each bike sharing stations in different areas of a city, they conclude this proposed model reach a better accuracy than the other deep and advanced models. This study might not contribute with our report as this is limit to basic machine learning methods. Karunanithi et al. [10] used advanced machine learning

techniques to forecast bike rental demand in urban areas, this study identify influential variables and ensure the development of resource-efficient and adaptable models, concluding the model which performs the best is Randon Forest with an accuracy of 77%, Karunanithi et. al. use similar features compared with our model such as temperature, wind speed, seasons, hours of the day and holidays that can support our result.

Feng and Wang [11] analyze a dataset similar that the one use for this report, with similar features, they use the conventional multiple linear regression model, however, the accuracy is too low, although a good linear relationship between factors and to forecast the multiple linear regression model is not applicable. For this reason, they include Random Forest model which increases significantly the accuracy to 82%.

Torres et. al. (2024) [12] included Random Forest to predict usage of bike-sharing where they expressed this model is an advisable option when a quick, simple prediction is required, on account of a new approach was considered, Artificial Neural Networks (ANN) this model provided the best results in terms of accuracy.

C. DATA MINING METHODOLOGY

To answer the research question "how do weather conditions and hourly usage patterns influence demand for bicycle rental systems in urban areas?" I used KDD (Knowledge Discovery in Databases), this methodology is a systematic process that seeks to identify useful data, this method was apply as follows:

Data selection: this dataset was chosen from UCI Machine Learning Repository and contains 13 columns and 17389 rows.

Data pre-processing: The columns 'instant' was removed as is just an index value, as well the column 'dteday' as we have 'year' and 'month' and finally the columns 'casual' and 'registered' are redundant as we just need the total count of bikes, this data is in column 'cnt'.

Missing values have not presented in the dataset, so any step was taken to handle them.

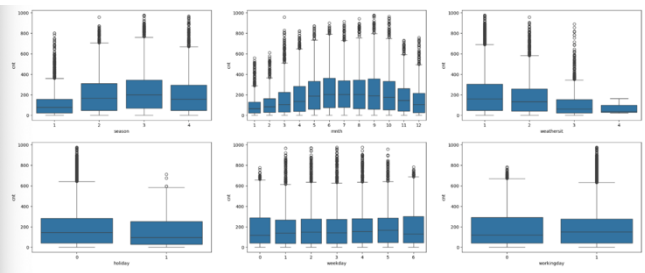


Figure 7. Boxplot of the categorical variables

From the boxplots shown above, we can observe many outliers present in these categories, in this case I decided to keep them considering important events during the year that affect the bike rental, justifications for every variable are the following:

- Season: Summertime is the season where people rented more bikes, has a trend this could be a good predictor.
- Month: this column has a trend for bookings, between June and September people rented more bikes, this could be a good predictor.
- Weathersit: When weather is clear, people rented more bikes, this could be a good predictor.
- Holiday: Mos of the bookings happened when is not holiday, this means data might be biased, this could be not a good predictor.
- Weekday: Weekday does not show any trend, this might impact the model.
- Workingday: Most of the bookings happened in a working day, this could be a good predictor.

The correlation matrix was analyzed as shown below:

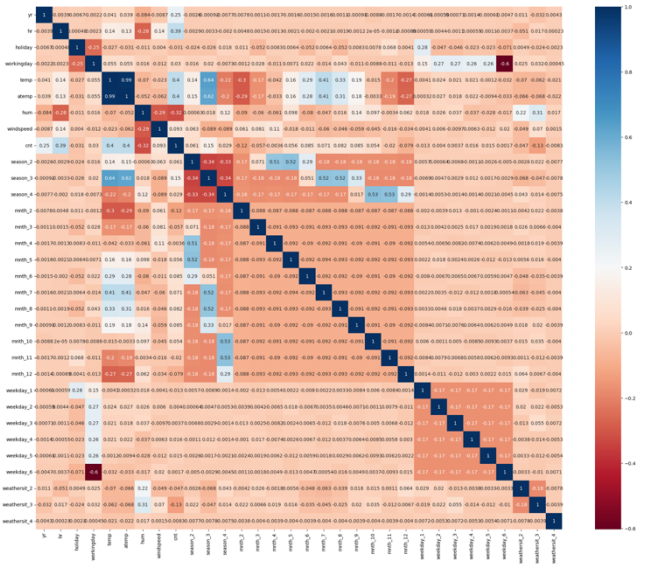


Figure 8. Correlation matrix

From the table above, correlation can be identified, we can see 'temp' and 'atemp' are highly correlated, so I will drop both 'atemp' and 'temp' as 'atemp' is just the feeling of the temperature and 'temp' as we have weather situation.

Data transformation: As we are dealing with a regression problem, dummy variables can be used for the categorical features, such as season, month, weekday and weathersit. To use dummy variables, it is necessary to convert the features into categorical data types and then create the dummy variables, dropping the first variable.

Data mining: Three machine learning methods (Linear Regression, Random Forest and Decision Tree) were analyzed for this dataset, the idea is to compare different measures, as the Mean Squared Error (MSE), Mean Absolute Error (MAE) and R squared, to select the best and effective method.

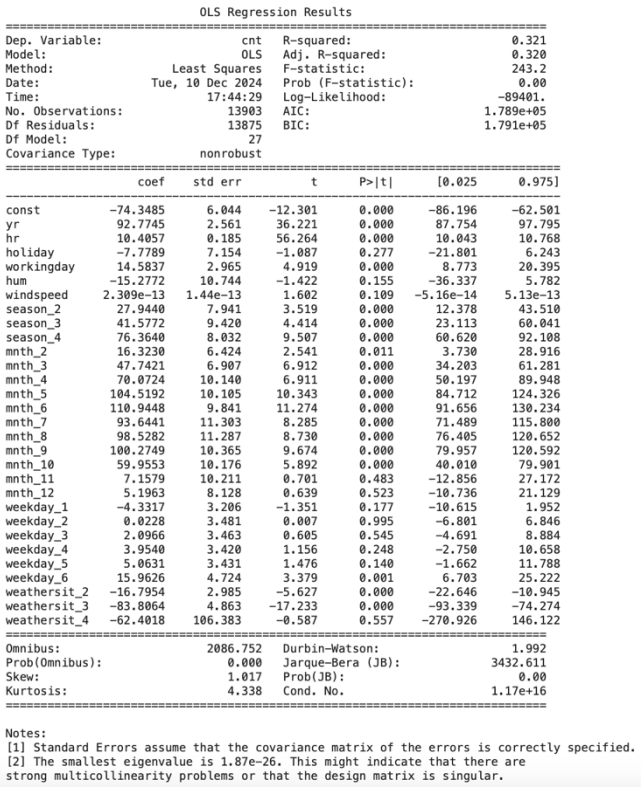
The dataset was split into training (80%) and test (20%) sets to assess how effectively our machine learning model works.

Data evaluation: This step is the last one used in KDD methodology, and it will be discussed in detail in section D.

D. EVALUATION

To answer the research question, " how do weather conditions and hourly usage patterns influence demand for bicycle rental systems in urban areas?", the machine learning models mentioned before were trained. Each model was evaluated using performance measures to ensure effectiveness and reliability. MAE measures the average of the residuals I the dataset, MSE measures the variance of the residuals, R squared is used to explain how well the independent variable explain the variability in the dependent variable. Using the KDD methodology, I focused on data mining and data evaluation to determine how effective is every model to predict loan approval.

1. Linear regression: was trained, and the summary model is shown below:



OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.321
Model:	OLS	Adj. R-squared:	0.320
Method:	Least Squares	F-statistic:	243.2
Date:	Tue, 10 Dec 2024	Prob (F-statistic):	0.00
Time:	17:44:29	Log-Likelihood:	-89401.
No. Observations:	13903	AIC:	1.789e+05
Df Residuals:	13875	BIC:	1.791e+05
Df Model:	27		
Covariance Type:	nonrobust		

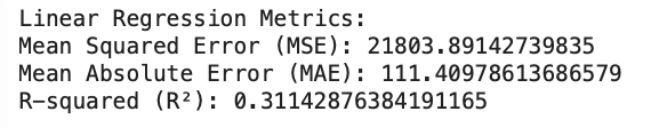
	coef	std err	t	P>|t|	[0.025	0.975]
const	-74.3485	6.044	-12.301	0.000	-86.196	-62.501
yr	92.7745	2.561	36.221	0.000	87.754	97.795
hr	10.4057	0.185	56.264	0.000	10.043	10.768
holiday	-7.7789	7.154	-1.087	0.277	-21.801	6.243
workingday	14.5837	2.965	4.919	0.000	8.773	20.395
hum	-15.2772	10.744	-1.422	0.155	-36.337	5.782
windspeed	2.309e-13	1.44e-13	1.602	0.109	-5.16e-14	5.13e-13
season_2	27.9440	7.941	3.519	0.000	12.378	43.510
season_3	41.5772	9.420	4.414	0.000	23.113	60.041
season_4	76.3640	8.032	9.507	0.000	60.620	92.108
mnth_2	16.3230	6.424	2.541	0.011	3.730	28.916
mnth_3	47.7421	6.907	6.912	0.000	34.203	61.281
mnth_4	70.0724	10.140	6.911	0.000	50.197	89.948
mnth_5	104.5192	10.105	10.343	0.000	84.712	124.326
mnth_6	110.9448	9.841	11.274	0.000	91.656	130.234
mnth_7	93.6441	11.303	8.285	0.000	71.489	115.800
mnth_8	98.5282	11.287	8.730	0.000	76.405	120.652
mnth_9	100.2749	10.365	9.674	0.000	79.957	120.592
mnth_10	59.9553	10.176	5.892	0.000	40.010	79.901
mnth_11	7.1579	10.211	0.701	0.483	-12.856	27.172
mnth_12	5.1963	8.128	0.639	0.523	-10.736	21.129
weekday_1	-4.3317	3.206	-1.351	0.177	-10.615	1.952
weekday_2	0.0228	3.481	0.007	0.995	-6.801	6.846
weekday_3	2.0966	3.463	0.605	0.545	-4.691	8.884
weekday_4	3.9540	3.420	1.156	0.248	-2.750	10.658
weekday_5	5.0631	3.431	1.476	0.140	-1.662	11.788
weekday_6	15.9626	4.724	3.379	0.001	6.703	25.222
weathersit_2	-16.7954	2.985	-5.627	0.000	-22.646	-10.945
weathersit_3	-83.8064	4.863	-17.233	0.000	-93.339	-74.274
weathersit_4	-62.4018	106.383	-0.587	0.557	-270.926	146.122

Omnibus:	2086.752	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3432.611
Skew:	1.017	Prob(JB):	0.00
Kurtosis:	4.338	Cond. No.	1.17e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.87e-26. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Figure 9. Linear Regression model

The following measures were done for the model:



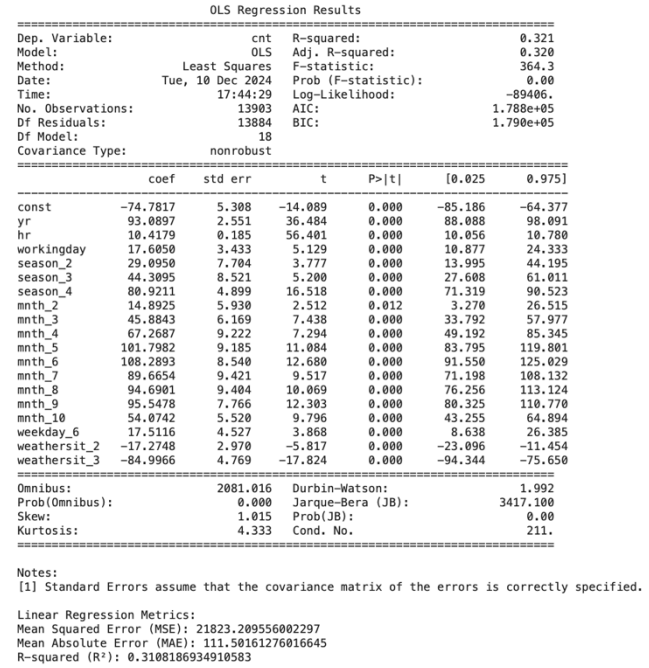
```
Linear Regression Metrics:
Mean Squared Error (MSE): 21803.89142739835
Mean Absolute Error (MAE): 111.40978613686579
R-squared (R²): 0.31142876384191165
```

Figure 10. Results for the Logistic Regression model

MSE: 21.823 indicates a high average error in the predictions, so this suggests the linear regression does not fit the data properly.

MAE: 111.5, this value is high, supports the MSE measure and the idea that the model does not fit the data well.
R2: 0.31, this value is low, as suggests that just the 31% of the variability of the data is explain for this model.

The summary results also shown a message at the end of the model suggesting strong multicollinearity, for this reason some columns with a high p-value were drop as they are not significant. The outcome after removing these columns is shown below:



OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.321
Model:	OLS	Adj. R-squared:	0.320
Method:	Least Squares	F-statistic:	364.3
Date:	Tue, 10 Dec 2024	Prob (F-statistic):	0.00
Time:	17:44:29	Log-Likelihood:	-89406.
No. Observations:	13903	AIC:	1.788e+05
Df Residuals:	13884	BIC:	1.790e+05
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-74.7817	5.308	-14.089	0.000	-85.186	-64.377
yr	93.0897	2.551	36.484	0.000	88.088	98.091
hr	10.4179	0.185	56.401	0.000	10.056	10.780
workingday	17.6050	3.433	5.129	0.000	10.877	24.333
season_2	29.0950	7.704	3.777	0.000	13.995	44.195
season_3	44.3095	8.521	5.200	0.000	27.608	61.011
season_4	80.9211	4.899	16.518	0.000	71.319	90.523
mnth_2	14.8925	5.930	2.512	0.012	3.270	26.515
mnth_3	45.8843	6.169	7.438	0.000	33.792	57.977
mnth_4	67.2687	9.222	7.294	0.000	49.192	85.345
mnth_5	101.7982	9.185	11.084	0.000	83.795	119.801
mnth_6	108.2893	8.540	12.680	0.000	91.550	125.029
mnth_7	89.6654	9.421	9.517	0.000	71.198	108.132
mnth_8	94.6901	9.404	10.069	0.000	76.256	113.124
mnth_9	95.5478	7.766	12.303	0.000	80.325	110.770
mnth_10	54.0742	5.520	9.796	0.000	43.255	64.894
weekday_6	17.5116	4.527	3.868	0.000	8.638	26.385
weathersit_2	-17.2748	2.970	-5.817	0.000	-23.096	-11.454
weathersit_3	-84.9966	4.769	-17.824	0.000	-94.344	-75.650

Omnibus:	2081.016	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3417.100
Skew:	1.015	Prob(JB):	0.00
Kurtosis:	4.333	Cond. No.	211.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Linear Regression Metrics:
Mean Squared Error (MSE): 21823.209556002297
Mean Absolute Error (MAE): 111.50161276016645
R-squared (R²): 0.3108186934910583

Figure 11. Logistic Regression model

As we can observe, the measures are the same.

2. Random Forest:

This model has a MSE (2.310) this indicates a lower error compared to Linear regression and it supports with the MAE (29.64) which is lower as well. R squared suggests just the 93% of the variability of the data is explain for this model.

3. Decision Tree:

This model has a MSE (3.98) this is a lower error, and the MAE (37.93) is consistent with the MSE. R squared suggests just the 87% of the variability of the data is explain for this model.

As we confirm from the literature revised, Random Forest performs well in this case of bike sharing services as performed in other reports, showing a high capacity and ability to predict. As we selected Random Forest, we evaluated feature importance, to see which variables are relevant for our model.

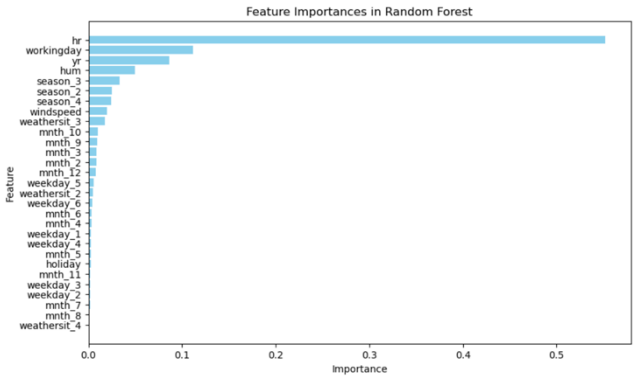


Figure 12. Feature importance

This help us to answer which factors are relevant when predicting bike sharing services. Hour and working day are the variables with the highest importance for the model.

4. Conclusions and Future Work

After modelling the different machine learning methods, the best model with the best measures is Random Forest (MSE: 2310, MAE: 29.64 and R-squared: 0.9270), the results suggests that Random Forest is the best model for this case, might be because this model managed non-linear relationships and capture complex patterns. The results answer partially our research question by demonstrating that hour is an important feature when predicting bike sharing services, however weather conditions might be not important.

Despite the good results, there are some limitations that are important to consider, given more time, I would include the variable temperature which I drop, and I would compare results, I would include deep learning methods to analyze the measures. These results might help bike rental companies to better scheduling for bike-sharing systems.

III. SECTION C: HOTEL REVIEWS ANALYSIS

***Abstract*—Analyzing and understanding the hotel reviews is essential to improve in this industry, giving important information to make decisions and implement better services. This study will use different techniques of NLP (Natural Language Processing) to analyze some reviews and identify critical patterns in the perception of the customers.**

The findings revealed that the three models used achieved a high accuracy. According to the performance measures, Random Forest and XGB reach a high value.

***Keywords*—hotel reviews, Random Forest, XGB regressor.**

A. Introduccion

Sentiment analysis helps to discover meaningful information about digital text, it helps to determine emotionality, positive and negatives sentiment in different fields. Comprehending hotel reviews is an essential key to giving feedback to the company and improving service. These reviews can cover different aspects such as guests experience, cleanliness, staff, amenities, food and overall service.

Customers always seek for trends and patterns across online reviews, it is important to understand the date of the reviews and keep in mind that those reviews are based on personal preferences and priorities. For instance, someone who is looking for facilities such as gym or a playground can be not interested in the buffet, however, other people might be interest in the buffet, so everyone will rate the hotel according to their necessities.

When going through a hotel review, people must consider that sometimes, some reviews are fake, and hotel owners pay people to post reviews. There are some key aspects to keep in mind when checking for reviews, for instance, check more than one review site (google maps, TripAdvisor, booking.com), do not trust in hotels with just few reviews, put more attention on recent reviews and check if the hotel gives any feedback in the review.

This report will find patterns in the reviews of the hotel to identify similar aspects and evaluate the sentiment expressed by customers. With this analysis we want to respond to what recurring themes and sentiments dominate hotel reviews, and how do these factors impact overall guest ratings.

B. Related Work

Hotel reviews is essential when people decide which hotel going on holidays, for this reason understand and analysis sentiment is an important key. Two of the big leaders in hotel bookings (booking.com and TripAdvisor) by Rita et al. [12] were compared in a study using ANOVA, among the results they conclude TripAdvisor's sentiment was always higher than in Bokking.com, these results are aligned with the outcome conducted here, as the hotels from the highest category were the ones with the highest average sentiment, with this similarity the study analyzed by Rita. et. al. contribute good to our research questions since some findings are the same.

Khanum et al. [13] analyzed sentiment on data gathered from the Twitter site. They created a model using different methods of machine learning and compared with a deep learning model using LSTM where this method obtained better accuracy in comparison with the machine learning models. This study was made for a different purpose however it is important to keep in mind that for sentiment analysis deep learning models might reach a better accuracy, for future work in hotel reviews analysis, deep learning methods might be built to compare the measures, and it is likely to reach a better result.

To answer our research question and understand which sentiments dominate hotel reviews, Nicolau et al. [14] conclude that prices increase with the level of sentiment and high process lead to low occupancy rate, although we cannot compare the accuracy between models due to Nicolau et. al. does not build a machine learning model; their conclusions support the analysis made for this report.

Interpreted which factors are relevant for customers when rating a hotel or writing a review is critical, this might depend on type of travel. Roy [15] found external factors like walking facility, external attraction and type of hotel influenced significative the reviews.

C. DATA MINING METHODOLOGY

To answer the research question "what recurring themes and sentiments dominate hotel reviews, and how do these factors impact overall guest ratings?" I used KDD (Knowledge Discovery in Databases), this methodology is a systematic process that seeks to identify useful data, this method was apply as follows:

Data selection: this dataset was selected from Kaggle and contains 17 columns and 515739 rows.

Data pre-processing: The columns used for this analysis were: 'negative review', 'positive review' and 'reviewer score' as they are important features for the analysis.

Any field contained missing values; however, we identified 507 duplicates. Those duplicates values were eliminated as they are a small position in the dataset.
Within this step a clean text process has been done to convert the texts in lower cases, delete numbers, special characters, punctuation and extra space. In the image shown below, the clean process can be observed:

	review	Reviewer_Score	\
488440	Would have appreciated a shop in the hotel th...	9.6	
274649	No tissue paper box was present at the roomNo...	8.8	
374688	Pillows Nice welcoming and service	7.9	
404352	No Negative Everything including the nice upgr...	10.0	
451596	No Negative Lovely hotel v welcoming staff	9.6	

	Clean_Review
488440	would have appreciated a shop in the hotel tha...
274649	no tissue paper box was present at the roomno ...
374688	pillows nice welcoming and service
404352	no negative everything including the nice upgr...
451596	no negative lovely hotel v welcoming staff

Figure 13. Sample of the clean data

In addition to the clean process, I used tokenization, which help me to split the text in small parts-tokens and lemmatization to use the word in the base form, for example, instead of 'walking' use 'walk'.
A list with some stop words were used to eliminate this words that cannot give a relevant meaning to the text, such as 'the', 'at', 'of', 'an', etc.
As a final step for this process, sentiment analysis was applied which give back a 'dictionary' with some key points: 'neg', negative sentiment score, 'neu', neutral sentiment score, 'pos' positive sentiment score and 'compound' the overall sentiment.

Data transformation: After completing the data cleaning, the text was transformed into numerical vectors, this helps to capture the semantic relationships between words and documents, and it is useful to build the models.
Term Frequency-Inverse Document Frequency (TF-IDF) was added it for every word to find out which word has useful information for the analysis. TF computes the number of times the word appears in the text and IDF computes the relative importance of every word, and this depends on how many times the word can be found.
To finish with the step, word clouds was made to visualize the sentiment analysis, I create positive and negative world clouds to identify the most representative words for positive and negatives sentiment. These figures are shown below:

Figure 14. Positive sentiment

Figure 15. Negative sentiment

Based on the figures, there is only slight difference between Positive Reviews and Negative Reviews. Both reviews are dominated with words like hotel, room and staff. But, in positive reviews there are also other words, that are much more positive and constructive, such as excellent, friendly and good. While negative reviews contain neutral to negative words, such as bad, poor, small and stay.
It means that even though both Positive and Negative Reviews mostly consist of the same words, but it has different meaning.

Data mining: Three machine learning methods (Random Forest, XGB Regressor and Ridge Regression were analyzed for this dataset, the idea is to compare different measures, as the accuracy, precision, recall, MSE and R squared to select the best and effective method.

The dataset was split into training (75%) and test (25%) sets to assess how effectively our machine learning model works.

Data evaluation: This step is the last one used in KDD methodology, and it will be discussed in detail in section D.

D. Evaluation

To answer the research question, " What recurring themes and sentiments dominate hotel reviews, and how do these factors impact overall guest ratings?", the machine learning models mentioned before were trained. Each model was evaluated using performance measures to ensure effectiveness and reliability. Accuracy measures overall performance, Precision measures positive prediction quality, Recall evaluates sensitivity to positive cases, F-1 score shows the balance between Precision and Recall, MSE measures the average squared difference between the predicted values and the actual values and R-squared measure the goodness of fit. Using the KDD methodology, I focused on data mining and data evaluation to determine how effective is every model to predict loan approval.

1. Random forest: was trained, and the summary model is shown below:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2186
1	1.00	1.00	1.00	10581
accuracy			1.00	12767
macro avg	1.00	1.00	1.00	12767
weighted avg	1.00	1.00	1.00	12767

Figure 16. Random Forest summary

This model is classifying all the classes perfectly, however, these results might be theoretical, as obtaining 100% of accuracy might be caused for overfitting or unbalanced data.

2. XGB Regressor:

The results for this model were, MSE: 3.916 * 10-6 and R-squared: 0.999.
This model is predicting high accurate values with errors that are not significant, R-squared can be an indicator of overfitting.

3. Ridge Regressor:

The results for this model were, MSE: 5.545 * 10-10 and R-squared: 0.999.
This model is predicting high accurate values with errors that are not significant, R-squared can be an indicator of overfitting.

E. Conclusions and Future Work

The results for the three models are highly accurate, it suggests that the model made a very well captured of the relationships between the features and the target variables in this case the ratings of the reviews, but in real world is unlikely to obtain similar outcomes. Having said that, it suggests overfitting, data leaking or evaluation issues as the training data can be similar to the test data.

However, the good results it is unlikely to conclude that any model might be used to answer our research question, whereas from the data analysis we can respond partially to the question, as from the word clouds we can observe words such as hotel, friendly, staff and helpful, this suggest that customers appreciate the service: and from the negative word cloud we can conclude that customers focused on issues related with rooms, could be the cleanliness, breakfast, could be the quality or variety. This helps us to identify that the service can lead to positive reviews and issues related with the facilities lead to negative reviews. In this case, new strategies can be focused on keep the good customer service and improve services and facilities.

For future work, cross validation can be used to measure the performance of the model and review the capacity to generalized to avoid overfitting.

References

[1] Consumer Finance Protection Bureau, "What is a credit score?" *Consumerfinance.gov*. Accessed: Dec. 13, 2024 [Online]. Available: https://www.consumerfinance.gov/ask-cfpb/what-is-a-credit-score-en-315/#:~:text=A%20credit%20score%20is%20a,information%20from%20your%20credit%20reports.

[2] B. Fay, "Defaulting on your loans," *Debt.org*. Accessed: Dec. 13, 2024 [Online]. Available: https://www.debt.org/credit/loans/default/.

[3] V. Sinap, "A comparative study of loan approval prediction using machine learning methods," *GU J. Sci., Part C*, vol. 12, no. 2, pp. 644–663, 2024. doi: 10.29109/gujsc.1455978

[4] N. Deborah, A. Rajiv, A. Vinora, C. Manjula, S. Mohammed, G.S. Mohammed, "An efficient loan approval status prediction using machine learning," , *in 2023 Int. Conf. on Adv. Comput. Technol. and Appl. (ICACTA)*, Oct. 2023. doi: 10.1109/ICACTA58201.2023.10392691

[5] N. Uddin, M. K. U. Ahamed, M. A. Uddin, M. M. Islam, M. A. Talukder and S. Aryal, "An ensemble machine learning based bank loan approval predictions system with a smart application," *Int. J. of Cogn. Comput. Eng.*, vol. 4, pp. 327–339, Jun. 2023. doi: 10.1016/j.ijcce.2023.09.001

[6] M. Madaan, A. Kumar, C. Keshri, R. Jain and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," in *IOP Conf. Ser.: Materials Sci Eng.*, vol. 1022, 012042, pp. 1–8, Oct. 2020. doi: 10.1088/1757-899X/1022/1/012042

[7] D. Dansana, S. G. K. Patro, B. K. Mishra and V. K. Prasad, "Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm," *Eng. Reports*, vol. 6, no. 2, e12707, Jun. 2023. doi: 10.1002/eng2.12707.

[8] European Cyclists' Federation, "Bike share schemes," *ecf.com*. Accessed: Dec. 13, 2024 [Online]. Available: https://ecf.com/what-we-do/urban-mobility/bike-share-schemes-bss

[9] Y. Pan, R. C. Zheng, J. Zhang, and X. Yao, "Predicting bike sharing demand using recurrent neural networks," *Procedia Comput. Sci.*, vol. 147, pp. 562-566, 2019. doi: 10.1016/j.procs.2019.01.217.

[10] M. Karunanithi, P. Chatasawapreeda and T. A. Khan, "A predictive analytics approach for forecasting bike rental demand," *Decision Anal. J.*, vol. 11, suppl. 1, 100482, May 2024. doi: 10.1016/j.dajour.2024.100482

[11] Y. Feng and S. Wang, "A forecast for bicycle rental demand based on Random Forest and multiple linear regression," in *2017 IEEE/ACIS 16th Int. Conf. on Comput. and Inf. Sci. (ICIS)*, pp. 101-105, May 2017. doi: 10.1109/ICIS.2017.7959977.

[12] P. Rita, R. Ramos, M. T. Borges-Tiago, and D. Rodrigues, "Impact of the rating system on sentiment and tone of voice: A Booking.com and TripAdvisor comparison study," *Int. J. of Hospitality Man.*, vol. 104, 103245, July 2022. doi: 10.1016/j.ijhm.2022.103245

[13] F. Khanum, P. S. Lakshmi, and H. V. Reddy, "Sentiment analysis using natural language processing, machine learning and deep learning,"

in *2024 5th Int. Conf. on Circuits, Control, Comm. and Comput. (I4C)*, Oct. 2024, pp. 113-118. doi: 10.1109/I4C62240.2024.10748425

[14] J. L. Nicolau, Z. Xiang, and D. Wang, "Daily online review sentiment and hotel performance," *Int. J. of Contemporary Hospitality Manag.*, vol. 36, no. 2, pp. 790-811, May 2023. doi: 10.1108/IJCHM-05-2022-0594

[15] G. Roy, "Travelers' online review on hotel performance – Analyzing facts with the Theory of Lodging and sentiment analysis," *Int. J. of Hospitality Manag.*, vol. 111, 103459, May 2023. doi: 10.1016/j.ijhm.2023.103459