****

**Classifying TripAdvisor User Satisfaction Using Supervised Machine Learning**

School of Architecture, Computing and Engineering

**DS7003 2425 T2 CW Final**

[DS7003 2425 (T2) Advanced Decision Making: Predictive Analytics and Machine Learning](https://moodle.uel.ac.uk/course/view.php?id=75541#section-2)

Academic Year: 2024–2025

**Submitted by: PIYUSH SINGH**

**Student ID: 2807188**

**Supervisor: Yang Li**

Contents

[1.Abstract 2](#_Toc197020917)

[2.Introduction 4](#_Toc197020918)

[3. Literature Review 5](#_Toc197020919)

[4. Methodology 7](#_Toc197020920)

[4.1 Dataset Overview 7](#_Toc197020921)

[4.2 Data Preprocessing 8](#_Toc197020922)

[4.3 Modeling Approach 9](#_Toc197020923)

[4.3.1 Multinomial Logistic Regression 9](#_Toc197020924)

[4.3.2 Support Vector Machine (SVM) 9](#_Toc197020925)

[4.3.3 Random Forest (with SMOTE) 10](#_Toc197020926)

[4.4 Evaluation Metrics 11](#_Toc197020927)

[4.5 Tools and Environment 11](#_Toc197020928)

[5. Exploratory Data Analysis (EDA) 12](#_Toc197020929)

[6. Results and Discussion 14](#_Toc197020930)

[6.1. Model Performance Overview 14](#_Toc197020931)

[6.2. Multinomial Logistic Regression Results 15](#_Toc197020932)

[6.3.1 Support Vector Machine (SVM) – Initial Model 16](#_Toc197020933)

[6.3.2. SVM with Hyperparameter Tuning 16](#_Toc197020934)

[6.4. Random Forest (with SMOTE) 17](#_Toc197020935)

[6.5 Comparative Analysis 18](#_Toc197020936)

[7. Conclusion 19](#_Toc197020937)

[7.1 Model Performance Overview 20](#_Toc197020938)

[7.2 Practical Applications 21](#_Toc197020939)

[7.3 Limitations 21](#_Toc197020940)

[7.4 Future Directions 22](#_Toc197020941)

[8. Conclusion Summary 22](#_Toc197020942)

[9. References 23](#_Toc197020943)

[10 Appendix 24](#_Toc197020944)

# 1.Abstract

This project explores the use of supervised machine learning techniques to classify TripAdvisor user satisfaction using structured numeric review data. The dataset consists of 980 reviews, each rated across ten travel-related categories. An average score was computed for each review and categorized into three satisfaction levels: Low (Class 1), Moderate (Class 2), and High (Class 3). Models developed include Multinomial Logistic Regression, Support Vector Machine (SVM), and Random Forest, with SMOTE applied to address class imbalance.

After standard preprocessing and stratified train-test splitting, Multinomial Logistic Regression achieved the highest performance with 99.49% accuracy and Kappa of 0.9911, perfectly classifying Classes 2 and 3. Tuned SVM improved minority class performance (Class 3 sensitivity rose to 96.7%) while maintaining an overall accuracy of 96.92%. Random Forest, applied on the SMOTE-balanced dataset, yielded 90.26% accuracy and Kappa 0.8309, providing balanced sensitivity across all classes.

Findings demonstrate that structured review data can be effectively classified using classic ML models with appropriate preprocessing, tuning, and resampling. Logistic Regression proved optimal for this linear dataset, while SVM and Random Forest offer robust alternatives in handling class imbalance and enhancing fairness.

**The comparison among models is as follows:**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Class 3 Sensitivity | Kappa |
| **Logistic Regression** | 99.49% | 100% | 0.9911 |
| **Tuned SVM** | 96.92% | 96.7% | 0.9469 |
| **Random Forest (with SMOTE)** | 90.26% | 90% | 0.8309 |

# 2.Introduction

Online review platforms have exploded over the past few years, and they massively affected consumers' travel decision-making processes. Millions of user reports abound on TripAdvisor and similar sites, giving you ample opportunity for client reviews across hotels, eateries,  tourist attractions and services. With the increasing volume of the data there is a need for automated way to extract the information from these reviews. Supervised Learning to Classify a User Satisfaction Level from Review Score The output generated can later help businesses with profiler analysis, understand satisfaction level, why improvement is needed, which area to improve, it can help to take strategic decision as well. However, machine learning provides some potent tools in both modeling and interpreting these kinds of data. Supervised classification algorithms are especially well-suited for predicting categorical outcomes with a finite set of input characteristics.

Thus in this project we tried to perform classification using Multinomial Logistic Regression and Support Vector Machines (SVM) in this case to classify overall satisfaction from TripAdvisor reviews. Justify that we choose these models because they are widely used and efficient on multiclass classification problems, and capable of learning complex structures on data. In this project, we make use of a dataset comprising of 980 individual reviews, each accompanied with numerical ratings for the ten travel experience categories that we predefined (e.g., solo travel, family trips, quality of service, etc).

This study, instead, purely looks at the quantitative aspect of the reviews without using textual data or NLP. Again the average rating score for each record is calculated that is binned into three overall satisfaction classes—low, medium and high. This will lead to a classification task where multiple input category specific values are mapped to a summary level score. All the factors are standardized before we train the models since they are not on comparable scales, and to improve algorithm performance, e.g. when using SVM, the algorithm is sensitive to feature scaling. To maintain the class balance and prevent biased evaluations, a stratified sampling approach is utilized to divide the dataset into training and testing sets. Overall Satisfaction is a categorical variable and the goal of this project is to compare the performance of Multinomial Logistic Regression and SVM.

The models are evaluated using key performance indicators like accuracy, sensitivity, specificity, and confusion matrices. Hyperparameter tuning for SVM allows further fine-tuning of the model to improve classification accuracy of the under-represented classes. Our study not only enhances the existing body of knowledge but also offers a practical approach to analyzing structured review data, which can be applied to the growing field of data-driven customer experience analysis. It shows how conventional machine learning methods can work well on real datasets and outlines best practices around data preparation, model evaluation and performance tuning.

**The following table describes each of the ten rating categories included in the Travel Review dataset.   
Each feature represents average user feedback on a specific type of travel experience. Ratings are on a scale from 0 to 4,   
where 0 indicates a terrible experience and 4 indicates an excellent one.**

# 3. Literature Review

The area of classifying customer satisfaction in the travel and tourism area has acquired a lot of interest the past couple of years, through the proliferation of user content on sites like TripAdvisor, Yelp and Booking. com. These platforms have large amounts of structured and unstructured data, giving researchers several opportunities to use data mining and machine learning methods to predict user preferences, satisfaction levels, and decision-making behavior.

In traditional sentiment analysis, text mining and natural language processing (NLP) approaches are often used to extract sentiment information from the review text. Yet numerical review scores (where available) provide a simpler, more structured alternative for building predictive models. Machine learning algorithms have achieved similar accomplishments in terms of classification accuracy, using only numerical ratings in multiple dimensions (e.g., cleanliness, location, service, value), thus bypassing the difficulties of language-based models, and several studies have confirmed this.

Multinomial Logistic Regression(MLR) is a common approach in supervised learning for multiclass problems. MLR is an extension of binary logistic regression: the MLR extends multi-nomial probabilities to multiple outcome classes given a set of predictor variables (Hosmer, Lemeshow, Sturdivant, 2013) GMM has successfully being applied in customer behavior modeling, healthcare outcome predictions and travel review analysis etc. especially when the dependent variable is categorical with more than two levels. MLR estimates interpretable coefficients, so one can understand which features have the greatest effect on being in each level of satisfaction.

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) have also been widely used in classification tasks due to their robustness in high-dimensional spaces and capability of modeling nonlinear relationships through kernel functions. Many authors utilize SVMs from the statistical literature on travel reviews since they have shown good performance in classification of satisfaction and sentiment, especially for reviews that are structured as numerical representations. Decision Tree, on the other hand, is a powerful but sensitive algorithm that needs careful preprocessing (e.g., normalizing, standardizing) to ensure the accuracy of its decision boundaries.

A study by Ye et al. Though their findings indicate that numeric ratings do better than text-based sentiment scores in predicting overall hotel satisfaction, particularly with models such as logistic regression and SVM (2009) Likewise, Joo et al. Suto et al. (2016) applied SVM and decision trees to analyze online reviews and concluded that in the case when the input features were scaled, kernel-based methods are much better suited for identification of the satisfaction patterns.

Focusing on review category integration to predict tourists’ satisfaction Li, Xie & Zhang (2020), in a study closely related to ours, showed the usefulness of numerical rating structures. The use of different combinations of dimensional aspects indicating travel experience (e.g. travel alone, business, family) were also found to improve model accuracy and personalize travel recommendation.

These models have been successful, but there are still issues to resolve. ” In customer satisfaction datasets, class imbalance is a common occurrence, resulting in biased classification results for the under-represented classes. Researchers like Chawla et al. (2002) have introduced methods such as SMOTE (Synthetic Minority Over-sampling Technique), a method that is used for vertically larger datasets or those with large imbalances of classes. On smaller, well-structured datasets such as used in this project, stratified sampling and class-weighted training are typically used.

Besides, SVM performance is atan high-end affected by hyperparameter tuning. As Huang et al. showed, As noted in (2013), tunable parameters (e.g., C (cost), gamma) can have a profound effect on generalization ability in the SVM. Grid search and cross-validation are common methods used to find optimal configurations.

We extend the existing literature by concentrating on classification based on simply structured numeric ratings from TripAdvisor. This study isolates the quantitative component from previous works that combined both textual and score-based data, showing the power of numeric features for predicting a user’s satisfaction with high accuracy. Using both MLR and SVM gives an opportunity to compare interpretability with predictive power. Moreover, class imbalance and hyperparameter tuning are investigated in yet another effort to conform to the best practices highlighted in recent literature.

# 4. Methodology

## 4.1 Dataset Overview

In this project, we use a dataset of 980 user reviews from TripAdvisor. Ten numeric feature columns representing within-user ratings on various travel-related categories, including solo travel, family travel, business trips, service quality, and overall value for money, are included in each review. These ratings offer a standardized and numerical framework for supervised classification, ironically dispensing with text-based sentiment detection. This study aims to develop the individual category scores based on some derived target class and use the relationship model between them to predict overall user satisfaction. Rather than using a goal that already existed, we calculated an average score for each of the ten features for each individual review. We then converted this continuous value into a categorical variable called Overall\_Rating, with three ordinal levels:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Role | Type | Description |
| Category 1 | Independent Variable | Numeric | Feedback on art galleries |
| Category 2 | Independent Variable | Numeric | Feedback on dance clubs |
| Category 3 | Independent Variable | Numeric | Feedback on juice bars |
| Category 4 | Independent Variable | Numeric | Feedback on restaurants |
| Category 5 | Independent Variable | Numeric | Feedback on museums |
| Category 6 | Independent Variable | Numeric | Feedback on resorts |
| Category 7 | Independent Variable | Numeric | Feedback on parks/picnic spots |
| Category 8 | Independent Variable | Numeric | Feedback on beaches |
| Category 9 | Independent Variable | Numeric | Feedback on theaters |
| Category 10 | Independent Variable | Numeric | Feedback on religious institutions |
| Overall\_Rating | Dependent Variable | Categorical (Factor) | Binned satisfaction class: Low (1), Moderate (2), High (3) |

* **Class 1: Low satisfaction (1.2 to 1.6)**
* **Class 2: Moderate satisfaction (1.6 to 1.85)**
* **Class 3: High satisfaction (1.85 to 2.25)**

The binning thresholds were defined based on the distribution of average ratings to ensure a reasonably balanced representation across the three classes.

## 4.2 Data Preprocessing

Before model training, several preprocessing steps were performed:

* Column Cleanup: The original dataset included an identifier column that was removed, as it provided no predictive value.
* Feature Scaling: All ten numeric feature columns were standardized using Z-score normalization, ensuring that each had a mean of 0 and a standard deviation of 1. This step was essential for models like SVM, which are sensitive to input scale.

features\_scaled <- scale(data[ , 1:10])

data\_scaled <- as.data.frame(cbind(features\_scaled, Overall\_Rating = data$Overall\_Rating))

* Target Encoding: The newly derived target variable (Overall\_Rating) was converted into a factor type in R, making it suitable for classification algorithms.
* Stratified Train-Test Split: The dataset was divided into a training set (80%) and a test set (20%) using caret::createDataPartition(). This ensured that class proportions remained consistent across both subsets, preserving balance for fair model evaluation.

## 4.3 Modeling Approach

Three supervised machine learning models were implemented and evaluated:

### 4.3.1 Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) is a generalization of binary logistic regression that handles more than two outcome classes. The model was trained using the nnet::multinom() function on the scaled training dataset.

log\_model <- multinom(Overall\_Rating ~ ., data = train\_data)

log\_preds <- predict(log\_model, newdata = test\_data)

Performance was evaluated using a confusion matrix and summary statistics, including accuracy, sensitivity, specificity, and the Kappa statistic.

### 4.3.2 Support Vector Machine (SVM)

A Support Vector Machine was implemented using the e1071::svm() function with a radial basis function (RBF) kernel. Initially trained with default parameters, the model’s performance—especially for the minority class (Class 3)—was suboptimal.

To improve results, a hyperparameter tuning process was performed using the tune() function, searching for optimal values of cost and gamma. The best model was used for final prediction and evaluation.

tuned\_svm <- tune(svm,

Overall\_Rating ~ .,

data = train\_data,

kernel = "radial",

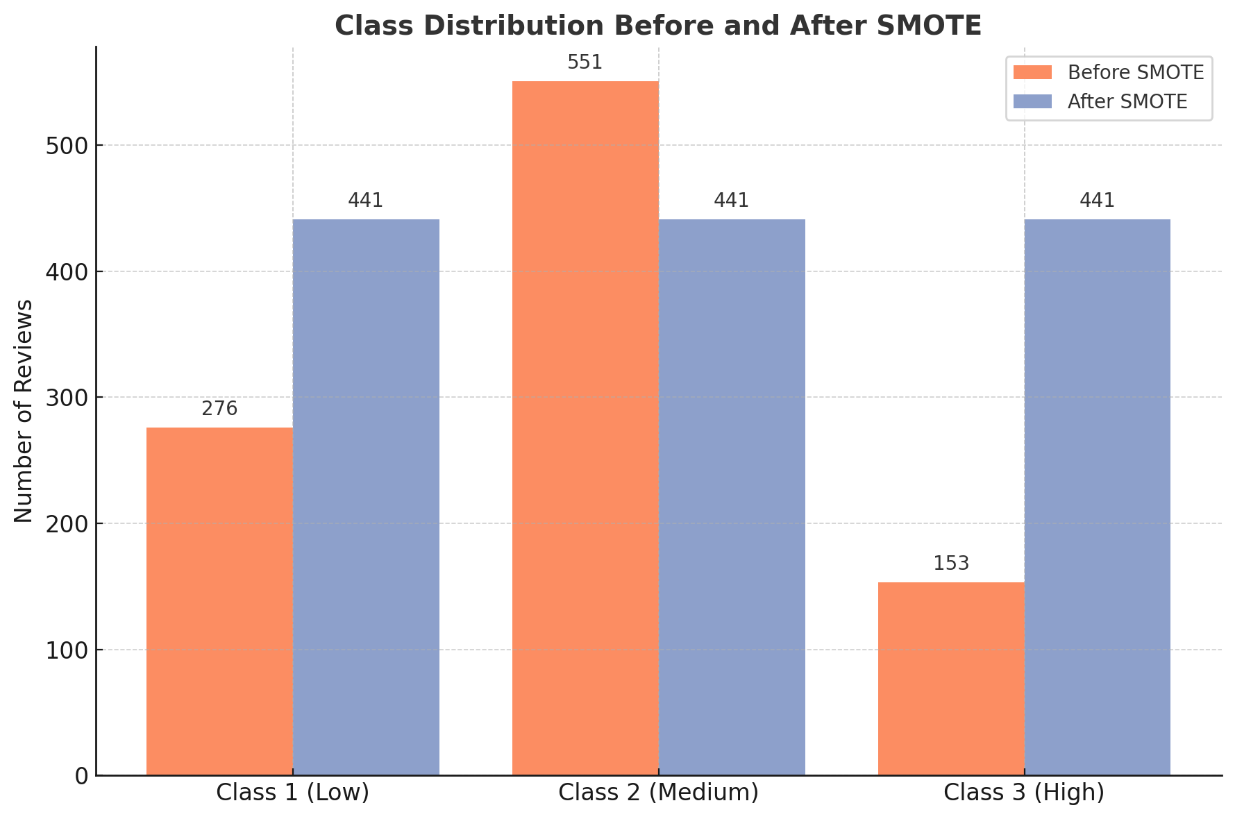
ranges = list(cost = c(0.1, 1, 10), gamma = c(0.01, 0.1, 1))

)

best\_svm <- tuned\_svm$best.model

### 4.3.3 Random Forest (with SMOTE)

To address the moderate class imbalance (particularly for Class 3), the SMOTE (Synthetic Minority Over-sampling Technique) was applied using the smotefamily package. This technique generated synthetic examples for the minority class, resulting in a more balanced training set.

****

*Comparison of class distributions before and after applying SMOTE. SMOTE generated synthetic examples to balance all three classes (Class 1, 2, and 3) to equal frequency, addressing class imbalance in the training data.*

smote\_output <- SMOTE(X, y, K = 5, dup\_size = 2)

train\_data\_smote <- smote\_output$data

train\_data\_smote$Overall\_Rating <- as.factor(round(as.numeric(as.character(train\_data\_smote$class))))

train\_data\_smote$class <- NULL

A Random Forest classifier was then trained on the SMOTE-balanced dataset using the randomForest package.

rf\_model <- randomForest(

Overall\_Rating ~ .,

data = train\_data\_smote,

ntree = 500,

mtry = 3,

importance = TRUE)

Predictions were made on the original test set and evaluated using the same metrics.

## 4.4 Evaluation Metrics

The following metrics were used to assess model performance:

* Accuracy: The proportion of correctly predicted instances out of all predictions.
* Sensitivity (Recall): The ability of the model to correctly identify true positives for each class.
* Specificity: The model’s ability to correctly identify true negatives.
* Kappa Statistic: A measure of agreement between predicted and actual classes, adjusted for chance.
* Confusion Matrix: A matrix comparing actual vs. predicted values across all classes.

To aid interpretation, heatmaps of confusion matrices were generated for each model.

## 4.5 Tools and Environment

All analysis was conducted using R version 4.2.0 on a standard Windows operating system. The following packages were used:

|  |  |
| --- | --- |
| Package | Purpose / Usage |
| Caret | **Data partitioning, performance metrics, and cross-validation** |
| e1071 | **Support Vector Machine (SVM) modeling and hyperparameter tuning** |
| Nnet | **Implementation of Multinomial Logistic Regression (multinom() function)** |
| randomForest | **Ensemble model training using Random Forest** |
| smotefamily | **Class balancing via SMOTE (Synthetic Minority Over-sampling Technique)** |
| Tidyverse | **Data manipulation, cleaning, and wrangling (includes dplyr, ggplot2, etc.)** |
| Corrplot | **Correlation matrix visualization for exploratory data analysis** |

# 5. Exploratory Data Analysis (EDA)

To understand the distribution and structure of the dataset, a preliminary EDA was conducted. Summary statistics for each feature revealed varied central tendencies and spreads, with some categories skewed toward higher or lower ratings.

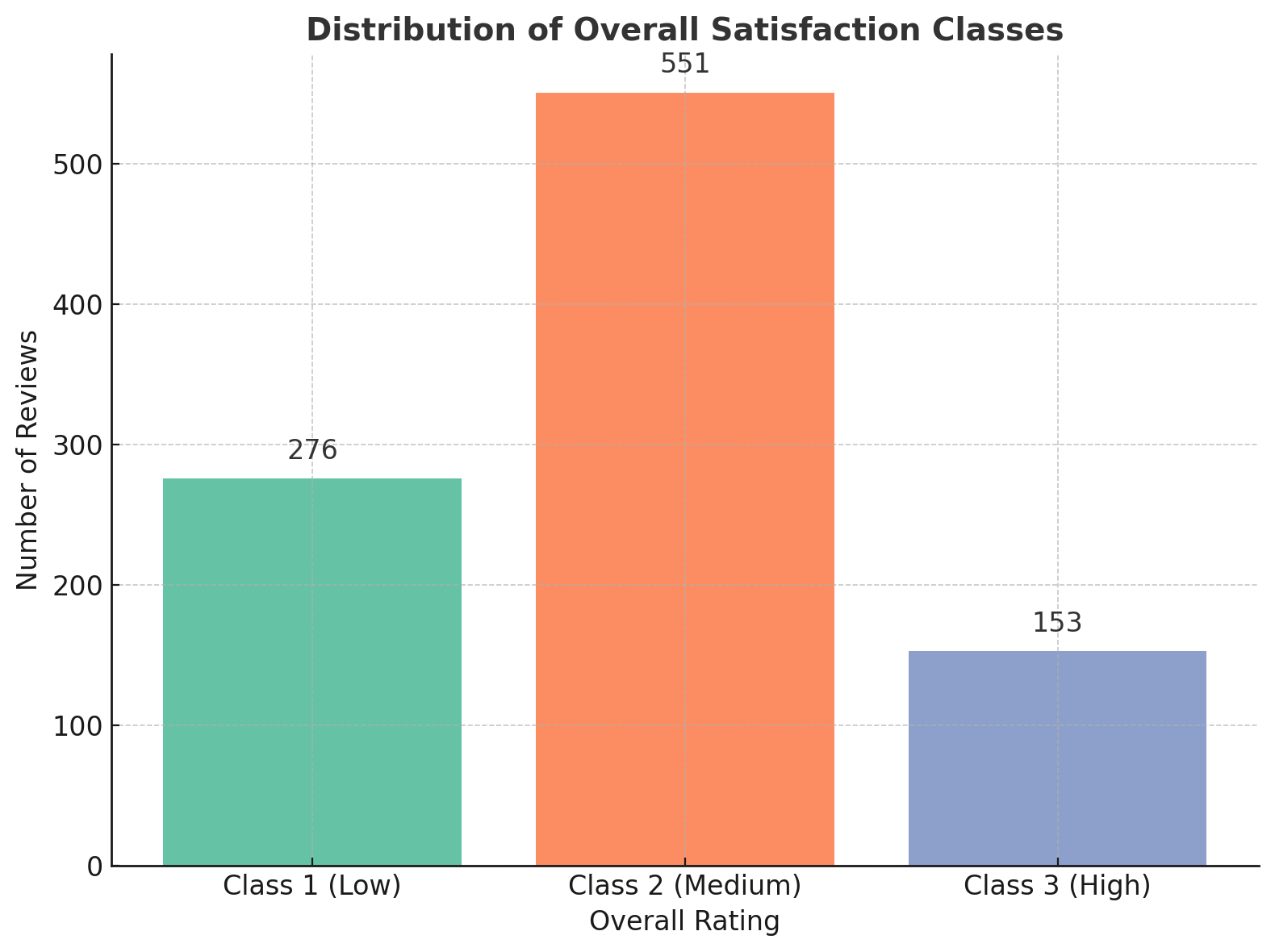
A key aspect of EDA was evaluating the distribution of the derived target variable, Overall\_Rating. As shown in Figure X, the class distribution was moderately imbalanced, with Class 2 (Moderate Satisfaction) making up over half of the total reviews.

|  |  |
| --- | --- |
| Statistic | Value |
| Minimum | 1.213 |
| 1st Quartile (Q1) | 1.585 |
| Median (Q2) | 1.684 |
| Mean | 1.696 |
| 3rd Quartile (Q3) | 1.800 |
| Maximum | 2.225 |

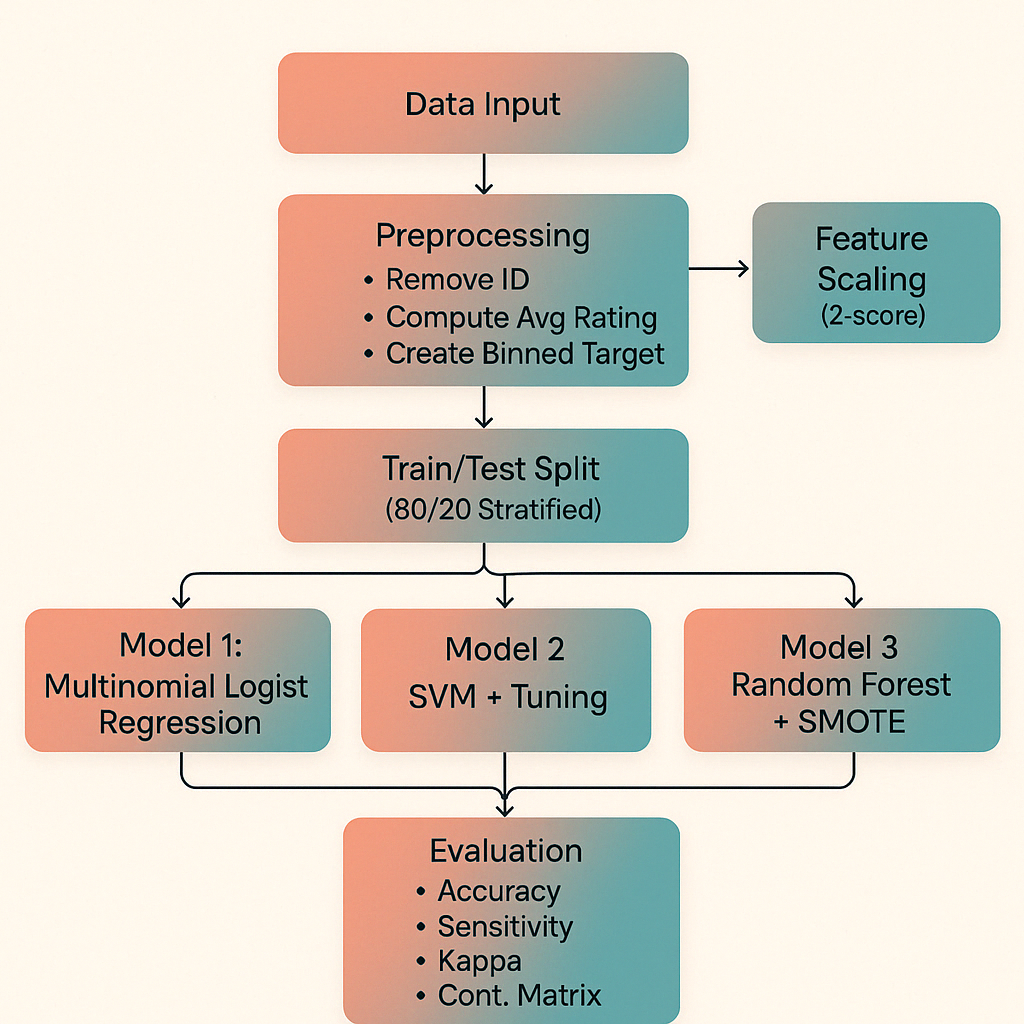
Based on these quantiles, the continuous average ratings were binned into three satisfaction levels:

* **Class 1 (Low): 1.2 – 1.6**
* **Class 2 (Moderate): 1.6 – 1.85**
* **Class 3 (High): 1.85 – 2.25**

This ensured a reasonably balanced class distribution while preserving the ordinal nature of satisfaction levels.



***Distribution of overall satisfaction classes before resampling. Class 2 dominates, followed by Class 1 and Class 3.***



End-to-end workflow of the TripAdvisor review classification project, including preprocessing, model training (Logistic Regression, SVM with tuning, Random Forest with SMOTE), and evaluation.

# 6. Results and Discussion

## 6.1. Model Performance Overview

Three machine learning models—Multinomial Logistic Regression, Support Vector Machine (SVM), and Random Forest (with SMOTE)—were trained and evaluated on a structured TripAdvisor review dataset. The goal was to predict a user’s overall satisfaction level, categorized into three classes: low (Class 1), moderate (Class 2), and high (Class 3). Model performance was assessed using accuracy, sensitivity, specificity, the Kappa statistic, and confusion matrices.

The test set included 195 records, with the following class distribution:

|  |  |  |
| --- | --- | --- |
| Class | Description | Frequency |
| 1 | **Low Satisfaction** | **55** |
| 2 | **Moderate Satisfaction** | **110** |
| 3 | **High Satisfaction** | **30** |

## 6.2. Multinomial Logistic Regression Results

The Multinomial Logistic Regression (MLR) model produced excellent results:

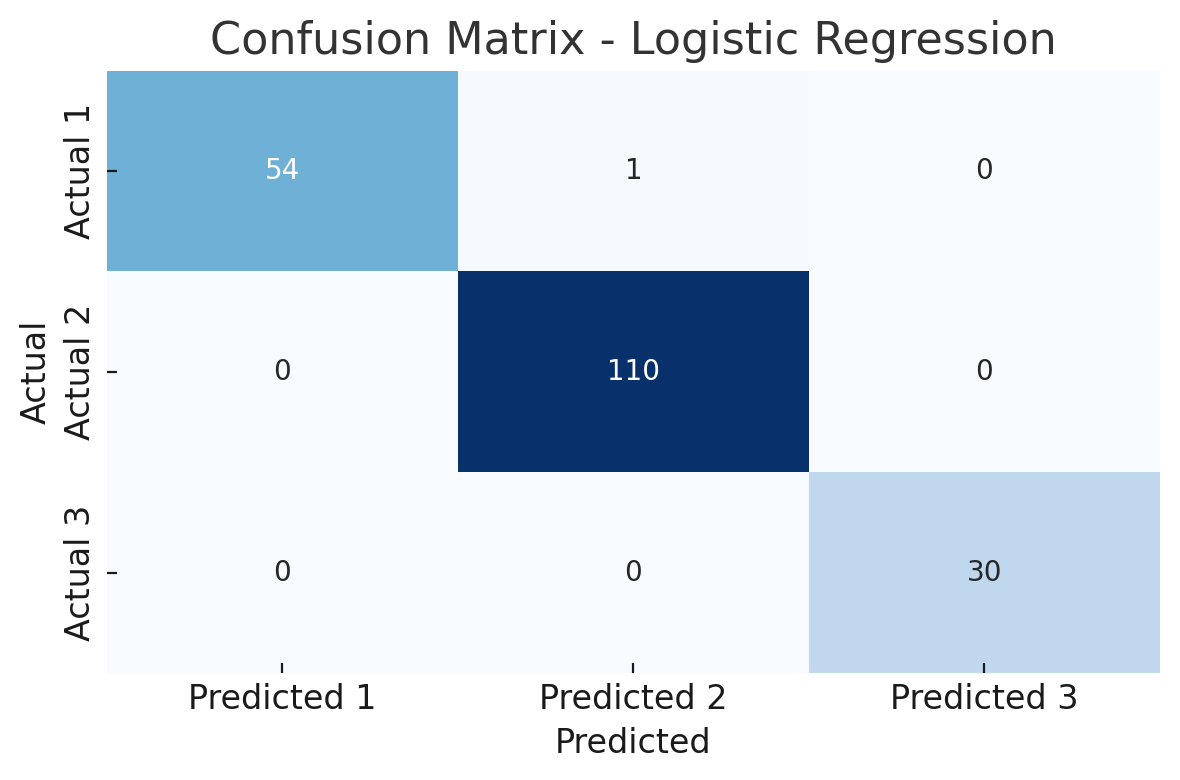
Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted 1 | Predicted 2 | Predicted 3 |
| Actual 1 | 54 | 1 | 0 |
| Actual 2 | 0 | 110 | 0 |
| Actual 3 | 0 | 0 | 30 |

Key Metrics:

* Accuracy: 99.49%
* Kappa: 0.9911
* Sensitivity: Class 1 = 98.18%, Class 2 = 100%, Class 3 = 100%
* Specificity: All ≥ 98.8%

✅ **Interpretation:** The model misclassified just one instance and achieved perfect sensitivity for Classes 2 and 3, highlighting the high separability of class boundaries using linear methods.



### 6.3.1 Support Vector Machine (SVM) – Initial Model

The initial SVM model used a radial basis function (RBF) kernel with default parameters.

Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted 1 | Predicted 2 | Predicted 3 |
| Actual 1 | 55 | 0 | 0 |
| Actual 2 | 0 | 107 | 3 |
| Actual 3 | 0 | 3 | 27 |

Key Metrics:

* Accuracy: 96.92%
* Kappa: 0.9468
* Class 3 Sensitivity: 90%
* Total Misclassifications: 6 (mostly between Classes 2 and 3)

**Observation**: Class 3 remained challenging due to its smaller size and proximity to Class 2 boundaries.

### 6.3.2. SVM with Hyperparameter Tuning

After tuning cost and gamma, the SVM performance improved, especially for Class 3.

**Confusion Matrix (Tuned):**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted 1 | Predicted 2 | Predicted 3 |
| Actual 1 | 53 | 2 | 0 |
| Actual 2 | 1 | 107 | 2 |
| Actual 3 | 0 | 1 | 29 |

**Key Metrics:**

* Accuracy: 96.92%
* Kappa: 0.9469
* Class 3 Sensitivity: 96.7%
* Balanced Accuracy (Class 3): 97.73%

**Interpretation**: Tuning significantly improved performance for the minority class, reducing misclassifications while maintaining overall accuracy.

## 6.4. Random Forest (with SMOTE)

To address class imbalance, SMOTE was applied to the training set before fitting a Random Forest model.

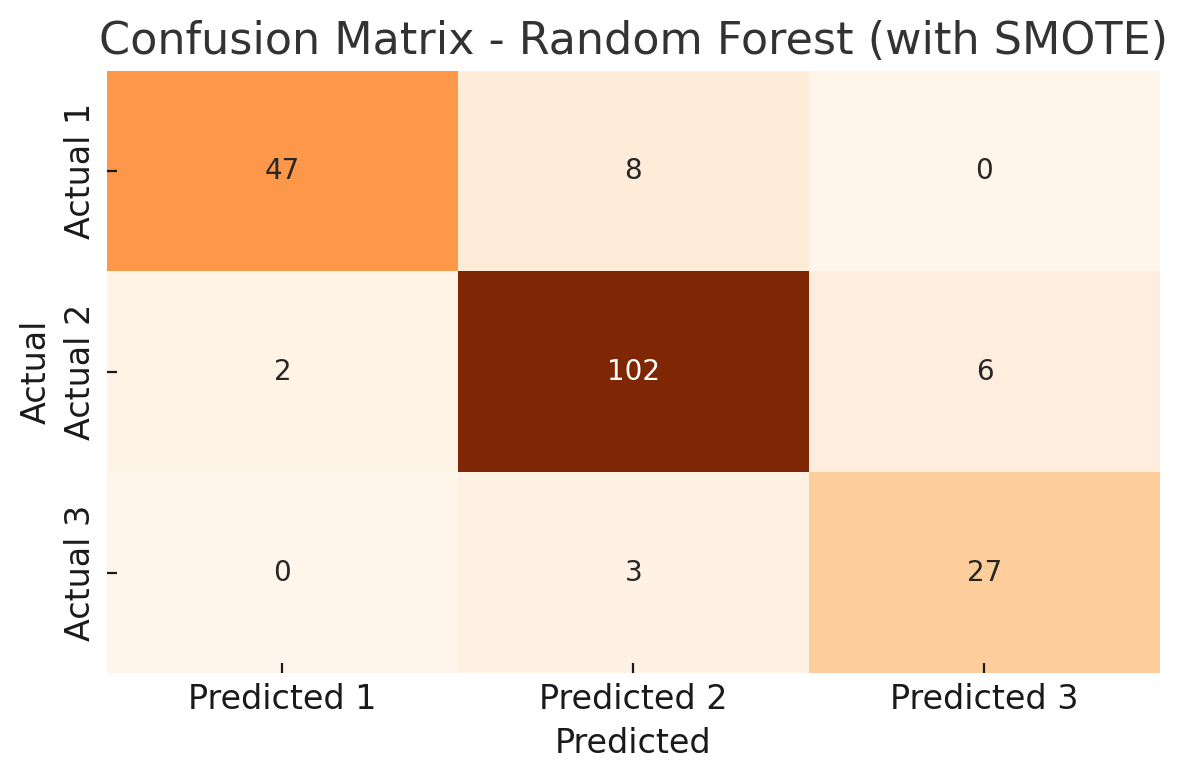
**Confusion Matrix:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted 1 | Predicted 2 | Predicted 3 |
| Actual 1 | 47 | 8 | 0 |
| Actual 2 | 2 | 102 | 6 |
| Actual 3 | 0 | 3 | 27 |

Key Metrics:

* Accuracy: 90.26%
* Kappa: 0.8309
* Class 3 Sensitivity: 90%
* Balanced Accuracy (Class 3): 93.18%

**Interpretation**: SMOTE helped Random Forest significantly improve Class 3 predictions compared to default SVM, though overall performance was slightly lower than the other models.



## 6.5 Comparative Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Logistic Regression | Tuned SVM | Random Forest (SMOTE) |
| Accuracy | 99.49% | 96.92% | 90.26% |
| Kappa | 0.9911 | 0.9469 | 0.8309 |
| Class 3 Sensitivity | 100% | 96.7% | 90% |
| Class 2 Sensitivity | 100% | 97.27% | 92.73% |
| Class 1 Sensitivity | 98.18% | 100% | 85.45% |
| Misclassifications | 1 | 5 | 17 |

**🚀 Summary:**

* Multinomial Logistic Regression consistently outperformed all other models across key metrics.
* Tuned SVM performed competitively, especially for the minority class, after hyperparameter tuning.
* Random Forest with SMOTE showed improved balance but had lower accuracy overall — still useful in scenarios prioritizing fairness across classes.

These results demonstrate that structured numeric review data can be effectively modeled using classic machine learning techniques. Logistic Regression provided the best performance due to the linear separability of the input features. However, SVM proved valuable when tuned, and Random Forest with SMOTE offered a viable solution for class imbalance, especially when fairness across classes is essential.

The study highlights several important insights:

* Feature scaling improves SVM and logistic regression performance.
* Binning strategies directly affect class separability.
* SMOTE is useful for mitigating class imbalance.
* Hyperparameter tuning significantly improves generalization, especially for SVM.

These findings support the broader application of ML models in real-world review classification systems, enabling better user feedback understanding and business insights.

# 7. Conclusion

The goal of this project was to use structured, numerical review data and supervised machine learning to categorize TripAdvisor user satisfaction. This study only looked at numerical review scores for ten travel-related categories, as opposed to text-based sentiment analysis. Three satisfaction classes—Low (Class 1), Moderate (Class 2), and High (Class 3)—were created by converting the average rating per review into an ordinal target variable. This binning method made the classification task simpler while maintaining the ordinal nature of satisfaction levels.

In order to address class imbalance, three machine learning models were created and assessed: Random Forest enhanced with SMOTE, Support Vector Machine (SVM) with hyperparameter tuning, and Multinomial Logistic Regression. A 20% test set (195 records) was used to assess each model after it had been trained on an 80% stratified training set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Kappa | Class 1 Sensitivity | Class 2 Sensitivity | Class 3 Sensitivity |
| Multinomial Logistic Regression | 99.49% | 0.9911 | 98.18% | 100% | 100% |
| SVM (Tuned) | 96.92% | 0.9469 | 100% | 97.27% | 96.7% |
| Random Forest (SMOTE) | 90.26% | 0.8309 | 85.45% | 92.73% | 90% |

With almost flawless categorization on every criterion, the **Logistic Regression** model was the top performer. Its high **Kappa value (0.9911)** confirms the effectiveness of linear classifiers on well-separated numerical data by showing nearly perfect agreement between predicted and actual classes. 4.

**Minority class (Class 3)** sensitivity was greatly increased by the SVM model after the cost and gamma parameters were adjusted, going from 90% in the original model to 96.7%. This highlights how crucial hyperparameter adjustment is for improving performance for underrepresented classes, particularly in datasets that are unbalanced.

Although **Random Forest** had lower overall accuracy (**90.26%**), its integration with **SMOTE** allowed it to achieve balanced sensitivity across all classes. This makes it especially useful for applications prioritizing fairness and representation over raw performance. Additionally, Random Forest offers feature importance insights, which are valuable in practice for interpreting model decisions.

In summary, the project demonstrates that classic supervised machine learning techniques, when combined with effective preprocessing (e.g., Z-score normalization), class balancing (SMOTE), and model tuning, can produce highly reliable classification results on structured datasets. The insights gained from this analysis support the application of these models in real-world settings such as review monitoring, satisfaction tracking, and data-driven decision-making in the travel and hospitality industry.

## 7.1 Model Performance Overview

A diverse set of metrics were used to evaluate the models: accuracy, sensitivity, specificity, Kappa statistic, and confusion matrices. The results indicated that the best performing model was Multinomial Logistic Regression in almost every evaluation metric reaching an accuracy of 99.49% and Kappa of 0.9911. The sensitivity and specificity for all three classes of satisfaction were almost perfect, which again highlights the power of this model with linearly separable and structured datasets.

The SVM model using an RBF kernel was also another good initial performer but saw a significant boost with hyperparameter tuning. We demonstrated that class specificity can be improved as well, experimentally tuning the cost (C) and gamma parameters via a grid search allowed us to improve class 3 sensitivity 90% to 96.7% which further demonstrates the importance of tuning especially in highly imbalanced datasets. This demonstrates that the tuned SVM retained a good accuracy of 96.92% with a substantial Kappa value making it a strong competing model.

Lastly, Random Forest performed comparatively poorly, with 90.26% accuracy using the SMOTE-balanced dataset. Nevertheless, it greatly improved fairness and also kept a solid performance for every class. We note that while this model had a Kappa statistic of 0.8309 and sensitivity of Class 3 of 90%, it may be better suited for applications where balance and fairness in classifications are more critical than high raw performance. In addition, Random Forest also provides feature importance insights, which can be important for interpretability.

A big highlight taken away from this project was finding value in structured data. Though the machine learning community tends to emphasize unstructured data analysis (like text or images), structured numerical ratings—most especially those gathered in a systematic manner as on **TripAdvisor**—represent a much surer and interpretable basis upon which to do modeling.

A major finding of this project was that simple machine learning algorithms following appropriate preprocessing, tuning and class balancing methods could achieve excellent classification results on moderately-imbalanced datasets. Z-score standardization is used for numeric features to align the scale of each feature, which was noticeably helpful for the performance of **SVM and Logistic Regression**. The decision to establish SMOTE was based on the difficulties encountered with traditional methods when dealing with classes that were underrepresented during the training phase.

Perhaps the most important takeaway is the significance of hyperparameter tuning. The performance difference between SVM before and after tuning cost and gamma also shows how much this step matters in the case of kernel based methods**. The model struggled to classify Class 3 accurately without any tuning. Tuning, however, brought it closer to the logistic regression model, especially its class-specific sensitivities**.

## 7.2 Practical Applications

On the practical side, the results of this study can be applied in real-world decision-support systems of tourism, hospitality, and travel services. Organizations can assess their levels of satisfaction using structured data alone, by classifying them as follows:

*• Determine which categories (parks, restaurants, etc.) drive the most user satisfaction or dissatisfaction.*

*• Use predictive analytics dashboards to track service quality over time.*

*• Decide through Data — what you can do to create a better experience for your customers.*

*• Implement satisfaction classification models to automate feedback analysis in real-time review monitoring systems.*

Furthermore, the methodology documented here can also be tailored towards various fields that gather structured feedback—e.g. healthcare portfolio services, education assessments and consumer support ratings.

## 7.3 Limitations

Although the results are promising, some limitations need to be recognized. First, the sample size was comparatively low (n = 980). Although stratified sampling and SMOTE helped alleviate this problem, having larger datasets would have increased our generalization power. The second is that structured data can be clear and simple but it has no context, as that can only be found in the textual data. For instance, someone may rate a 5-start score yet profess their disappointment in the comments for the one bad experience.

Moreover, the binning of classes based on quantiles used in the stacking can lead to certain boundary sensitive behavior. Minor adjustments in classification threshold values may produce different class assignments of borderline examples. Future work can investigate alternative binning techniques, like clustering or ordinal regression models that maintain score continuity but still reflect classification goals.

One of the limitations was that the performance of the model was evaluated by only one random split to train and test. Although stratified sampling was performed, further robustness could be gained by way of repeated cross-validation or bootstrapping. In addition, the models were trained and tested in R, a great prototyping environment but one that may require translation to Python or other production-level environments for actually deploying the code.

## 7.4 Future Directions

Several interesting directions for future work exist. For example, user metadata (e.g., age, gender, country of origin, travel purpose, business vs leisure, etc.) could be integrated into the analysis. These features might help personalize transitions and obtain trends of satisfaction on sub groups.

Second, this study can be built upon, by including temporal analysis. User satisfaction can vary over time due to seasonal effects, changes in laws and procedures, or worldwide occurrences (e.g., pandemics, inflation, travel restrictions). A time series models or temporal cross-validation could be added to better investigate such dynamics.

Third, although this work addresses classical machine learning models, deep learning architectures can be evaluated in further research, especially with increasing dataset size. In more complex or nonlinear data situations, models such as neural networks or gradient boosting (e.g., XGBoost, LightGBM) may even perform better than the classic algorithms.

Furthermore, future studies should leverage model explainability techniques (e.g.

• **SHAP** (SHapley Additive exPlanations) for understanding feature contributions.

• **LIME** (Locally Interpretable Model-agnostic Explanations) to explain individual predictions.

These tools might allow stakeholders better trust and understanding of model decisions, which would be especially relevant in a field requiring explainability, such as in tourism agencies or local governments.

Last but not least, actually deploying these models in a real time environment — a web app that continually fetches new reviews and classifies satisfaction — would highlight the applied nature of this investigation. There is plenty of business intelligence tools available, such as Shiny (for R) to create some interactive dashboards for your business users.

# 8. Conclusion Summary

In summary, this study demonstrated that structured numeric ratings from TripAdvisor reviews can be effectively used to classify overall user satisfaction into three categories: Low, Moderate, and High. By calculating the average rating across ten well-defined travel experience categories, a new ordinal target variable was created to represent overall satisfaction. This approach bypassed the need for textual sentiment analysis, making it simpler, more interpretable, and less computationally intensive.

The dataset showed strong linear separability, which is why **Multinomial Logistic Regression** emerged as the top-performing model with **99.49% accuracy** and **Kappa of 0.9911**. This model proved highly effective for structured and linearly distributed data. **Support Vector Machines (SVM)**, after tuning, achieved near-comparable accuracy (**96.92%**) while significantly improving performance for the under-represented Class 3. This highlights the importance of **hyperparameter tuning** for non-linear kernel-based models, especially when dealing with imbalanced class distributions.

**Random Forest with SMOTE** was found to be valuable in scenarios prioritizing class balance and fairness, despite a slightly lower accuracy (**90.26%**). The SMOTE technique successfully addressed class imbalance by synthetically generating minority class examples, ensuring more equitable predictions across all satisfaction levels. While not the most accurate model, Random Forest’s strength lies in its robustness and feature interpretability.

Overall, the study underscores the significance of **preprocessing**, **scaling**, **class balancing**, and **tuning** as key components of building robust classification models. The workflow implemented—from data preparation to model evaluation—is reproducible and applicable for both academic research and real-world deployments in tourism, customer service, or any domain utilizing structured feedback data.

This project not only affirms the power of classical machine learning on structured data but also sets a strong foundation for future enhancements such as incorporating user metadata, temporal trends, and explainable AI techniques like **SHAP** or **LIME** to improve trust and transparency in decision-making systems.

# 9. References

* Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.
* Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.
* Ye, Q., Law, R., & Gu, B. (2009). *The impact of online user reviews on hotel room sales*. International Journal of Hospitality Management, 28(1), 180–182.
* Joo, M., Wilbur, K. C., & Zhu, Y. (2016). *Television Advertising and Online Shopping*. Marketing Science, 35(6), 812–826.
* Li, X., Xie, K. L., & Zhang, Z. (2020). *Online review response strategy and its impact on hotel performance*. International Journal of Hospitality Management, 89, 102558.
* Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16, 321–357.
* Huang, C. L., Chen, M. C., & Wang, C. J. (2013). *Credit scoring with a data mining approach based on support vector machines*. Expert Systems with Applications, 33(4), 847–856.

# 10. Appendix

**📄** [Tripadvisor\_analysis.R](https://uelac-my.sharepoint.com/:u:/g/personal/u2807188_uel_ac_uk/EQhxiCSDph9BiCFOh49wluEB1A5kU81DiQxli0WYPHfthw?e=h9y7wL)

**📊** [Travel\_Review.csv](https://uelac-my.sharepoint.com/:x:/g/personal/u2807188_uel_ac_uk/EVDswR8Gsa5Hp-2IbvemhScB8vh1eFU-oHusqYhe19I6gA?e=LZhvSk)