**CHAPTER 5: CONCLUSION AND FUTURE WORK**

**CONCLUSION**

In this research, I have developed machine-learning algorithms on oversampled and under-sampled data separately to predict the customer who might cancel the reservation based on the customer details provided and the best performance model has been chosen on the performance comparison of models on both the sampled data individually and also, the data has been analyzed through the data visualizations to identify the possible reasons behind the cancellations. The Data understanding is crucial for this project as an initial step to understand the significance of each column. In the pre-processing steps, checked for the missing values and none of them were found. But there are a few columns that might not make any impact on the data, those two columns have been removed namely, the ID and the type of meal plan. There are 3 categorical columns, two of them are independent variables and one is a target variable, the one-hot encoding is applied to the independent variables namely “room\_type\_reserved” and “market\_segment\_type” which created 12 columns for the categories in it. In the next step, data visualizations are created to analyze all the columns of the data. Cancellation status by room type reserved, and market segment type are analyzed and the average price per room and lead time are plotted with histogram and some important findings are observed and noted in the results section. Counts of booking status have shown great imbalance in the dataset, hence, to deal with it, data is separated into the train, validation, and test sets in the data separation section. Later the sampling techniques operation is understood in the literature review, and they have been implemented separately for all the algorithms on both over-sampled and under-sampled data. Applying the SMOTE technique for oversampling the actual training data for the minority class, resulted in 41584 rows and 26 columns for independent data and 41584 rows for target variables, later This data was standardized and used to train the machine learning models with the best parameters obtained from the grid search on every algorithm namely, KNN, Decision Tree and Random Forest and the results are tabulated. Similarly, the RUS technique for under-sampling the actual training data to reduce the majority class, which resulted in the independent data with shape (20082,26) and (20082, ) for the target variable. Under-sampled data is standardized and fed to the machine-learning algorithms and results are tabulated and analyzed, below are the observations made from the data visualizations and from the model performances on sampled data.

From the data visualizations, it is observed that 38% of the lead time column data is falling between the gap range of 90 to 450 days which can be the important reason for cancellations people might get other plans to do or change their mind and also average price per room shows the people mostly preferred to book the rooms with price between 60 to 100 euros. Also, peak times for most bookings are August, September, and October with most of them adults with no children preferring to stay 1 or 2 weekend nights or 1 to 3 weeknights. Room type 1 is the most chosen with 25% of cancellations, and most people preferred to book online, people who booked online have taken advantage of canceling them, and it accounts for 25% as well.

The machine algorithms trained and tested on oversampled data have achieved the maximum performance on the random forest algorithm with 99.46% training accuracy, 90.55% validation accuracy, and 89.99% test accuracy on unseen data. The decision tree stands in the second position with 99.46%, 87.08%, and 88.33% for train, validation, and test accuracy. KNN stands 3rd and logistic regression is the base model with 77.68% accuracy on test data.

On under-sampling data, Random Forest had given the highest training, validation, and testing accuracy with 97.2%, 89.34%, and 89.25%. Similar to oversampled results, the decision tree stands in second position with 85.21% on test data, KNN stands 3rd with 84.38% on test data and the logistic regression model is the least performant model with only 78.23%.

In both the cases of under-sampled and over-sampled, Random feature importance has resulted in ‘lead time” and “average price per room” as are most important columns in predicting the hotel reservation cancellation of the customer provided the details. All the research questions have been answered with the help of a literature review and experimentation results.

On a conclusion note, “the null hypothesis has been rejected” as machine learning algorithms trained on oversampled data have achieved the highest performance results compared to the model's performance on under-sampled data. The best machine learning model for predicting hotel reservation cancellation is the random forest trained on over-sampled data with a 90% confidence rate on unseen data.

**FUTURE WORK**

There are many algorithms, like neural networks, support vector machines, K-means clustering, etc., that can be used to experiment on this hotel reservation data. much advanced architectures like artificial neural networks with learning parameters and loss functions can give more generalized performance and better accuracy. Also, alternative approaches to deal with imbalanced data have to be explored. These are things I have observed to consider for future work.