Classification Models for Hotel Booking

In this project, I would like to make a classification system for relevant hotel booking information. The occupancy rate is critical to the hotel's operation, and it is of great significance to forecast and analyze the occupancy rate. Through predictive analysis of occupancy rate, hotel managers can develop strategies to increase occupancy rate or adjust the booking process to save costs and increase revenue. In this project, I will use different classification models to determine which model is more suitable for predicting occupancy rate. The classification models I would like to evaluate are the Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, Naive Bayes, and Support Vector Machines.

First of all, the hotel booking dataset contains a lot of information, including hotel information, whether the customer canceled the booking, the types of the customer, whether the customer is a repeat customer, and the arrival time of the customer, etc. This dataset has a lot of columns representing various factors that may affect the hotel occupancy rate. Since there are too many factors, I decide to retain some factors that I think may have a greater impact on the occupancy rate, such as customer type and market segment. As can be seen from the results of the correlation plot, lead time has the greatest relationship with the booking cancellation rate. Then, we can learn a lot of useful information from the visualization results. The pie chart shows that there are four different types of customers in this dataset, namely transient, group, contract, and transient party customers. As a result, **The most frequent customer type is transient which has the largest part in the pie chart and the least frequent customer type is the group.** In the count plots, we can see that the cancellation rate of booking is quite high. This is beyond my imagination, as the count plot shows that the number of customers who canceled bookings is more than half of those who do not cancel bookings. Also, we can see there are only a few customers are repeat customers. I think **if the hotel can increase more repeat customers, it will increase profits**. The results of the distribution plot show that most customers booked hotels online.

After cleaning the data, I converted the market segment, deposit type, and customer type columns into numerical values which a mathematical model can process. I would like to know if these factors can affect whether the customer wants to cancel the booking. Next, I split the data into the test data and train data, and then fit the training data into different classification models. **In this dataset, features (X) are all the factors that may affect whether the customer will cancel the booking (target, Y).** I set all columns except the "is_canceled" column to features and set the "is_canceled" column to target. After successfully fitting all models into training data, I used 5-fold cross-validation to evaluate the accuracy rate of each model. According to the results, the cross-validation accuracy of the Logistic Regression model is 77.06%. The cross-validation accuracy of the Decision Tree model and the Random Forest model is 76.43%, and 77.90%, respectively. The cross-validation accuracy for the KNN model is 77.90% and for the Naive Bayes model is 55.19%. The cross-validation accuracy for linear, radial basis function and polynomial kernel in the Support Vector Machines model is 76.84%. The sigmoid kernel in the Support Vector Machines model has a cross-validation accuracy of 70.04%. Therefore, **the Random Forest model and KNN model had the highest cross-validation accuracy, while the Naive Bayes model had the lowest cross-validation accuracy.**

In addition, I generated a classification report for each model and put the results in one graph (figure 1) and one data table (figure 2). The graph can help me observe the differences between each model more clearly and intuitively. As can be seen from the graph, **all metrics of**

the Naive Bayes model are lower than other models, but the precision of the Naive Bayes model looks the same as other models. The sigmoid kernel metrics results appear to be the lowest of all the Support Vector Machines kernels. From the graph results, all models look similar, except for the obvious differences in Naive Bayes model metrics. Therefore, the data table can help me understand the gap between different models in more detail. In the classification report, the accuracy of the Logistic Regression model, the Decision Tree model, and the Random Forest model is 77.0%, 76.0%, and 78.0%, respectively. The accuracy for the KNN model is 77.0%, and for the Naive Bayes model is 53.0%. The accuracy for linear and polynomial kernels in the Support Vector Machines model is 76.0%. The accuracy of the radial basis function kernel is 78.0%, while that of the sigmoid kernel is 69.0%. **The accuracy of models in the classification report is similar to that of cross-validation accuracy. However, KNN, Naive Bayes models, linear, and polynomial kernels showed lower accuracy in classification reports than 5-fold cross-validation accuracy**.

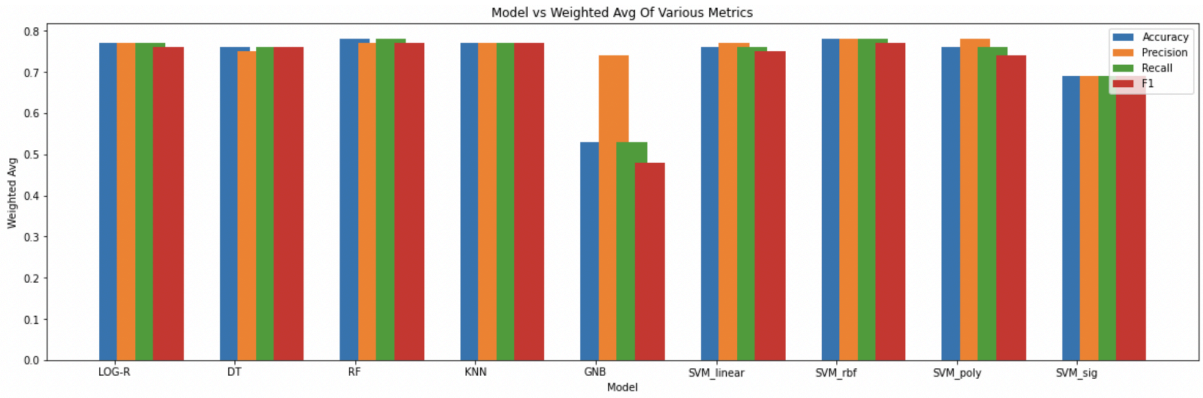
Moreover, the precision result for Logistic Regression, Random Forest, KNN model, and the linear kernel in the Support Vector Machines model is 77.0%. The precision of the Decision Tree model is 75%, and the precision of the Naive Bayes model is 74.0%. The precision of the polynomial and radial basis function kernel in the Support Vector Machines model is 78.0%. The sigmoid kernel's Precision is 69.0%. **The results show that polynomial and radial basis function kernels in the Support Vector Machines model have the highest precision.** In addition, the recall score for Logistic Regression is the same as the KNN model, which is 77%. The recall score for the Decision Tree model is the same as the linear and polynomial kernel in the Support Vector Machines model, which is 76%. The recall score for the Random Forest model is the same as the radial basis function kernel, which is 78%. The recall score for the Naive Bayes model is 53% and for the sigmoid kernel is 69%. **The results showed that the Random Forest model and radial basis function kernel had the highest recall score, while the Naive Bayes model had the lowest recall score.** Furthermore, the Logistic Regression and Decision Tree model have an F1 score of 76%. Random Forest and the KNN model have an F1 score of 77%. The F1 score for the Naive Bayes model is 48%. In the Support Vector Machines model, the F1 score of the linear kernel is 75%, the F1 score of the radial basis function kernel is 77%, and the F1 score of the polynomial kernel is 74%, and The sigmoid kernel's F1 score is 69%. **Random Forest and the KNN model have the highest F1 score.**

According to the results, the Naive Bayes model is the worst among all models. All metrics of this model are far lower than other models except for its precision. The Random Forest model has the best performance in cross-validation accuracy, accuracy in classification report, precision, recall score, and the F1 score. The radial basis function kernel in Support Vector Machines is the best performer with higher metrics than the other kernels, while the sigmoid kernel is the worst performer. In general, if hotel managers want to use the classification model to predict whether users are likely to cancel the booking, they can choose the Random Forest model to predict. The Random Forest model is easier to handle larger datasets. Managers should not choose a Naive Bayes model, which assumes independent predictors. There are so many predictors in this dataset that it may not be possible to produce a completely independent set of predictors.

In conclusion, the accuracy of all models is close to 80% but no higher than 80% except for the Naive Bayes model and sigmoid kernel which are much lower than 80%. I think we can improve the accuracy of these models in a few different ways. First of all, the dataset in this project is very large, which contains too many factors. So, I think we can improve the accuracy

of the model by processing the data. For example, we need to better understand different features and improve the accuracy of the model by selecting features that are more relevant to the target. We can also adjust the K value in the K-fold Cross-Validation to determine the allocation of Cross-Validation that better matches the validity of the evaluation model. If the classification model is more accurate, hotel managers will be more accurate at predicting hotel occupancy.

Appendix



(figure 1)

	Model	CV Accuracy	Accuracy	Precision	Recall	F1
0	LOG-R	77.06	77.0	77.0	77.0	76.0
1	DT	76.50	76.0	75.0	76.0	76.0
2	RF	77.90	78.0	77.0	78.0	77.0
3	KNN	77.90	77.0	77.0	77.0	77.0
4	GNB	55.19	53.0	74.0	53.0	48.0
5	SVM_linear	76.84	76.0	77.0	76.0	75.0
6	SVM_rbf	76.84	78.0	78.0	78.0	77.0
7	SVM_poly	76.84	76.0	78.0	76.0	74.0
8	SVM_sig	70.04	69.0	69.0	69.0	69.0

(figure 2)