Applied Analytics Project

Aanchal Chadda

Chapman University

MGSC 310

Professor Tim Frenzel

December 17, 2023

PROJECT OVERVIEW

The topic I chose to focus on for my applied analytics project was hotel cancellation rates. I wanted to look into the various factors of a hotel booking to understand which factors proved to be the most significant in predicting whether a hotel reservation is canceled or not. A recent report from Skift Research in June 2023 reports a Travel Health Index score of 103 relative to a baseline of 100, which shows us that the global travel industry has completely rebounded. In order to maintain the post-Covid highs the travel industry is currently experiencing, it's important to focus on the traveler experience. Whether a guest's hotel reservations are canceled or not is, arguably, the greatest indicator of a trip being confirmed. Through Exploratory Data Analysis (EDA) and the machine learning models I've employed, I can investigate further into the linear and non relationships between a variety of factors to see if a hotel booking is canceled or not. This information allows me to help hotel owners & management understand the booking behavior of their guests to ensure that cancellation rates remain at an all time low.

DATA SOURCE & PREPARATION

My dataset is sourced from Kaggle, and consists of 36,275 entries of hotel guests that made an initial hotel reservation (done pre-Covid). There were 19 unique variables that describe the hotel guest, the time of their booking (weekday vs weekend), whether or not they had children, whether or not they required parking, whether or not they were a past guest, and whether or not they canceled a reservation. To begin my data cleaning process, I first utilized the complete cases function to ensure no entries would have any missing values. In addition, I removed the Booking ID variable because it's unique to each guest and it doesn't contribute to providing any valuable insights in predicting the booking status. Lastly, I converted the remaining necessary categorical features of my data (meal plan, booking status, meal plan, room-type reserved, and market segment type) into dummy variables which created new numerical columns. This way I could remove the original columns that featured the same data in a categorical form, as data needs to be entirely numerical before I build my machine learning model.

MODEL SELECTION & RATIONALE

Originally, I was planning on utilizing a Ridge regression model for this problem. However, I decided against it because Ridge is most useful for problems with a smaller sample size. Additionally, ridge reduces coefficients whereas I was hoping to look at the coefficients to determine the most significant predictors of the booking status. Therefore, I chose my initial model to be a logistic regression model. Logistic regression models are

also utilized frequently for problems concerning binary classification (where the y/target variable is 0 or 1). Since I'm predicting whether or not a reservation will be canceled (1 for yes, 0 for no), I believe that a logistic model is an ideal one to begin with. I also tested out a decision tree, which proves to be useful in classification problems and provides an easy-to-interpret visual representation of the model. The final model I utilized was a Random Forest model, which utilizes multiple decision trees to output multiple predictions.

FEATURE ENGINEERING & SELECTION

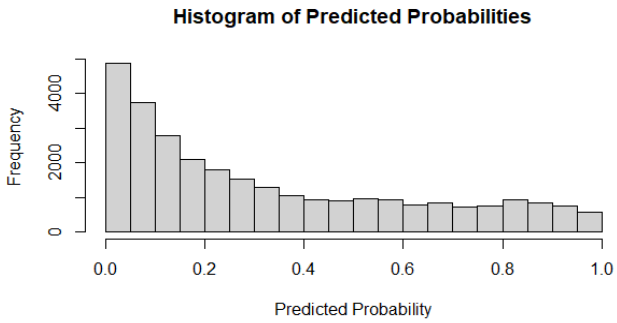
With my first model, I had to revise it by making sure I removed spaces in column names so that I wouldn't get errors when calculating the accuracy. Doing this improved the accuracy of my initial logistic model from 0.55 to 0.57, even after my logistic model got to work. In terms of my random forest model, I chose to prevent overfitting by pruning my decision tree prior to running the model. This ensures that the model itself is also not simply replicating the outcomes from the training data. I chose to include every variable from the start in my random forest variable because I believed that it would provide a more holistic review of every numerical & categorical feature in my booking.

MODEL TRAINING & VALIDATION

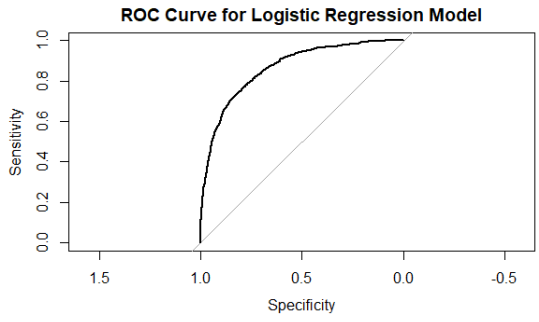
To train my model, I conducted a traditional 80/20 split between my testing and training data & gave a threshold of 0.5. I also set seed before splitting my data to ensure reproducibility in my results. The main metrics I utilized to validate my data are the accuracy, coefficients, & ROC-AUC. I also built an unpruned and pruned decision tree. The main reason I started with an unpruned tree was to ensure any complex patterns were captured in my training data prior to testing it on unseen (testing) data. The main drawback of an unpruned decision tree is that it's prone to overfitting, which means the model might get a little too comfortable with the training data and focus on replicating the performance of that, rather than focusing on being reliable with new data. Which leads to creating a pruned decision tree, which is known to make the model more simple & easily readable, making it less prone to overfitting (ideal for testing data). Since I'm hoping to find out what the greatest predictors are for booking status, I also chose to visualize my pruned decision tree to see how the model operates on test data. Additionally, with my random forest model, I tried to ensure the accuracy results were somewhat reproducible by changing the parameters to see if the accuracy rates would be similar in each forest. All of these models are best utilized for predicting cancellation rates as they are suited for

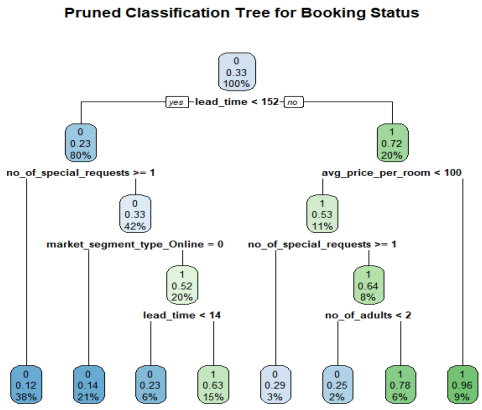
classification problems because they all have a “probabilistic nature" as my professor calls it.

RESULTS & INTERPRETATION



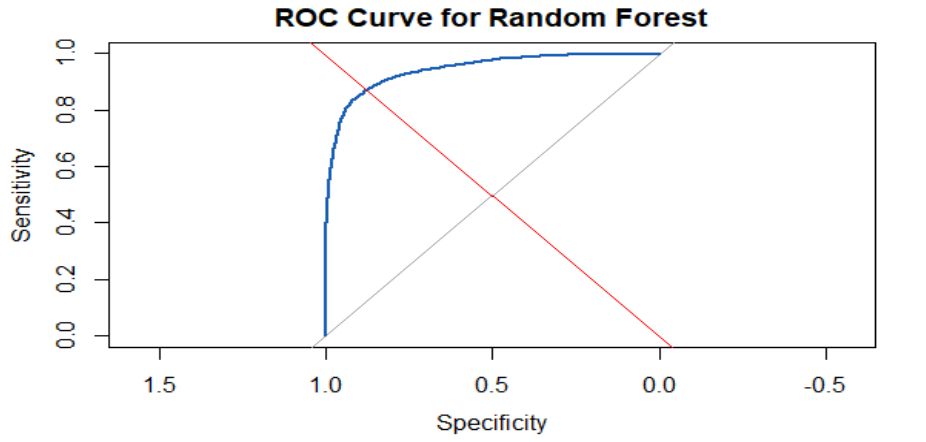
This histogram from the logistic regression model shows that the model predicts more cancellations, as it is skewed heavily towards 0. The logistic regression model provided an extremely low AIC (Akaic Information Criterion), which supports the conclusion that this initial model may prove to be a good fit based on its performance. The low p-values and high z-values of lead time, required car parking space, average price per room, and number of special requests suggest that these variables are the most important contributors to cancellation rates. For a hotel, this may mean a lower average price per room, increased car parking space, reasonable accommodation for special requests (which means a client has more interest in their booking), and a lower lead time (as higher lead times make it more convenient for a customer to make a cancellation as less penalties are applied. In terms of the ROC curve, the bow of the curve is going towards the top left which tells me that the model is good at differentiating between a positive and negative result to make an accurate prediction.





In terms of my classification tree (above), what is most significant is the lead time as that variable is at the root of the tree. With a lead time of under 152, if there’s 0 special requests, there’s an 80% chance the booking is cancelled. If there’s 1 or more special requests, and the booking is not made online (market segment variable) there’s a 42%

chance the booking is canceled, then seeing if lead time is under 14 to then evaluate average price per room. On the right, it looks at average price per room and adults.



The Random Forest algorithm provides an AUC of 0.943, suggesting the model is effective in differentiating between the binaries of 1 and 0. Forest 1 (F1) has the highest accuracy of 0.892, followed by F2 with 0.892 and F3 with 0.867. I used the default parameter, increased trees, and then increased variables with each split for the forests. As a hotel manager, I would utilize F1 since it has the highest accuracy.

LESSONS LEARNED & CHALLENGES

One of the main lessons I learned during this project was that looking at accuracy primarily isn't the best way to evaluate the performance of a model. For example, with many of my coefficients in my logistic regression, if I utilize the z-test the Pr(>|z|) value is exactly 1 which could indicate that the model is extremely saturated with many predictors that actually add zero value to my model. In addition, since there were so many iterations, maybe my model would benefit from more simplicity which would be achieved if I eliminated some variables and then ran the model. In terms of my random forest, I'm quite impressed with the AUC of 0.943, as the accuracy rates of all forests were pretty high so a higher AUC at least backs up the accuracy.

FUTURE WORK & IMPROVEMENTS

The first thing I wish I focused more on was delving deeper into the AUC ROC values. To expand on this project, I hope to look into the Gini coefficient as a higher Gini means the model is more reliable. I also think looking at the F1 score would prove to be a valuable statistic to show to shareholders, as F1 not only values precision but also takes into account the recall values as it measures the mean of both (recall and precision). It would also be helpful if I added a confusion matrix to show a visual representation of my

model's accuracy even in an initial model. The second thing I wish I focused on utilizing in my model would be the mean decrease gini, which essentially would tell me, after removing a variable, how much a model's accuracy decreases by. This would improve my analysis on the overall accuracy of a model as I'd have a more detailed evaluation of each variable. In addition, even though I'm new to machine learning, I wish I tried out the XGboost (extreme gradient boosting) algorithm. XGboost is designed to run cross-validation every iteration (for example there were 16 iterations in my logistic regression model so cross-validation would be run 16 times). It is easily scalable so it can handle large datasets like mine, works with imbalanced datasets, and can handle data in all sorts of industries (finance, healthcare, marketing, etc). Additionally, since I sourced my dataset from Kaggle, I looked at some of the past projects of competition winners and found that most of them utilized XGboost in their models (even for non classification problems). Lastly, I think it would've been cool if I utilized PCA (Principal Component Analysis) to visualize the summary of my findings as PCA would transform all of my variables into principal components, and would help in the clustering process that I would want to show through a visual. All of these improvements would improve the quality of my initial analysis, communicate my findings effectively, and help the hotel industry at large by finding the most accurate predictors of cancellation.