Mark Hoyt
03/06/24
DSE 6211
Final Project

Executive Summary:

ABC Hotels is trying to identify bookings with a high risk of cancellation. We have a variable of 0 or 1. 0 being the customer did not cancel and 1 being the customer did cancel, which gives us our dependent variable of booking_status. This model will help ABC hotels understand how other independent variables influence the likelihood of a customer canceling their reservation. I will be comparing two Feedforward Dense Neural Networks to process the data and predict whether or not a customer is likely to cancel. I am using this type of model because it works well with regression and classification-supervised learning. Another form of processing I will be conducting is the standardization of columns and removal of outliers which will be identified in the data processing stage of this project. Dates will be converted into a 'Seasons' variable as I hypothesize with my knowledge of this industry that the season may influence whether a customer cancels their reservation. The columns I will be including in my models are; type_of_meal_plan, room_type_reserved, and arrival_date. I believe with the correct model ABC Hotels will be able to accurately predict the likelihood of a customer canceling based on historical customer data. Customers will be clustered based on the data and how it matches up with similar historical data for customers. Two models were tested, a basic three-layer neural network and a more complex neural network with 4 layers and regularization.

Approach & Data:

1. I split the data into a training and test set

```{r}
#Training set data
training_set <- data[training_ind, ]

#Test set data
test_set <- data[-training_ind, ]
```

2. Changed booking status into a binary where 0 represents a cancellation and 1 is not canceled

```
#replacing Booking status to 0 and 1
training_set$booking_status <- ifelse(training_set$booking_status == 'canceled', 0, 1)
test_set$booking_status <- ifelse(test_set$booking_status == 'canceled', 0, 1)
```
```
```

3. Identified the variable that I believed had an influence on whether a booking was canceled or not (type_of_meal_plan, room_type_reserved, arrival_date)
4. Changed 'arrival_date' into 'season' as I believe the season a room is booked could influence the likelihood of cancelation

```
```{r}
# Assign the season to the new column
training_set$season[month %in% c(1, 2, 12)] <- "Winter"
training_set$season[month %in% c(3, 4)] <- "Spring"
training_set$season[month %in% c(5, 6, 7, 8)] <- "Summer"
training_set$season[month %in% c(9, 10, 11)] <- "Fall"
```
```

5. Summarized 'type_of_meal_plan' and identified 'meal_plan_3' only had 4 instances in our data set. So I combined it with 'not_selected' and turn this variable into 'meal_plan_other'

A tibble: 4 x 2

training_set$type_of_meal_plan <chr>	count <int>
meal_plan_1	20900
meal_plan_2	2444
meal_plan_3	4
not_selected	3831

4 rows

6. Summarized room_type_reserved and identified room_type 2, 3, 5, 6, and 7 had significantly less instances the room_type 1 and 4. I combined those room_types into 'room_type_other'

A tibble: 7 x 2

training_set$room_type_reserved <chr>	count <int>
room_type1	21074
room_type2	529
room_type3	5
room_type4	4538
room_type5	191
room_type6	729
room_type7	113

7 rows

7. Summarized ‘market_segment_type’ and identified ‘aviation’ and ‘complementary’ had fewer instances then the other segments so I combined them into ‘other_segment’

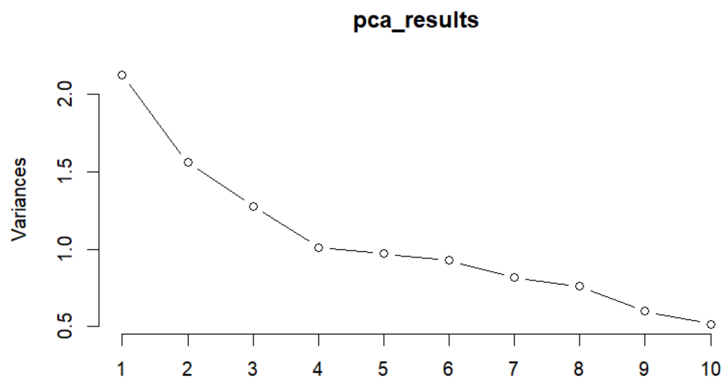
training_set$market_segment_type <chr>	count <int>
aviation	100
complementary	281
corporate	1471
offline	7846
online	17481

5 rows

8. I removed all other variables from the data set in order to prep for one hot encode
9. I scaled all the data

```
mean <- apply(training_set, 2, mean)
sd <- apply(training_set, 2, sd)
scaled_training_set_features <- scale(training_set, center = mean, scale = sd)
scaled_test_set_features <- scale(test_set, center = mean, scale = sd)
```

10. Conducted one hot encode on the data
11. Performed standardization on the data to ensure equal evaluation
12. Applied PCA

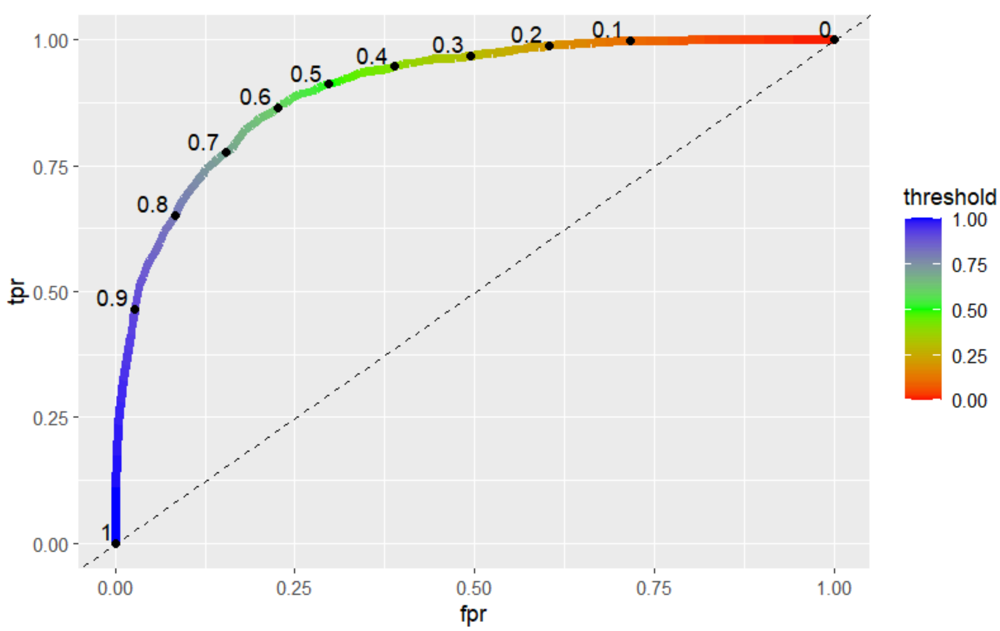


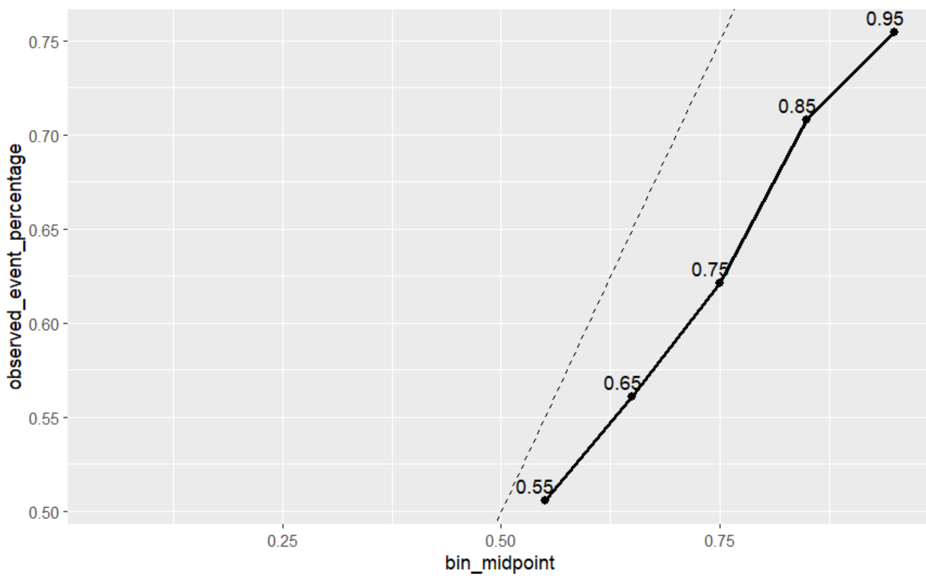
13. Utilized two dense neural network models

Two Dense Neural Networks were used and compared on our data. Both networks had ‘ReLU’ activation functions, ‘rmsprop’ optimization, and ‘binary_crossentropy’ loss function

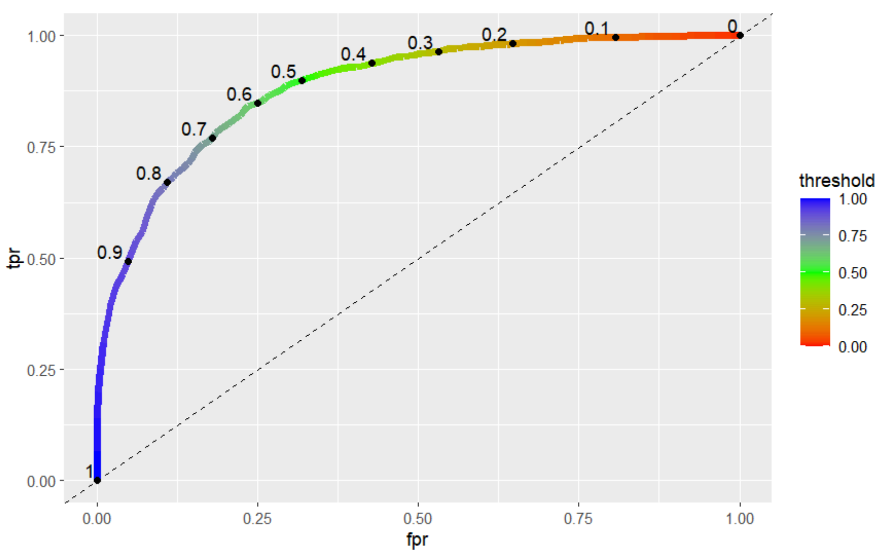
Model 1 had three layers, with layer 1 consisting of 100 units, layer 2 50 units, and layer 3 with 1 unit.
Model 2 had four layers, layer 1 consisting of 100 units, layer 2 50 units, layer 3 25 units, and Layer 4 with 1 unit. I also implemented Regularization to aid the overfitting and increase the model's effectiveness

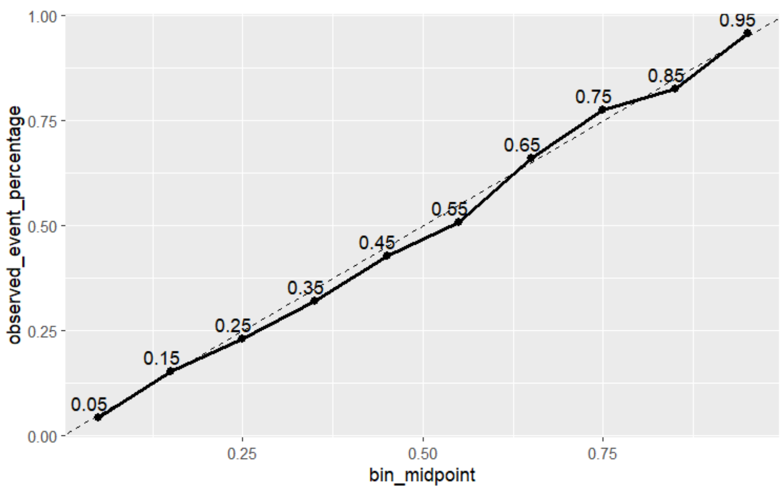
Model 1 had 50 epochs and resulted in an AUC score of 0.9051 with a promising ROC curve. And a calibration curve that resulted in over fitting





Model 2 also had 50 epochs but 4 layers which resulted in an AUC score of 0.8856. This model calibration curve fits very closely to the line and only briefly falls to underfitting





Detailed Findings and Eval:

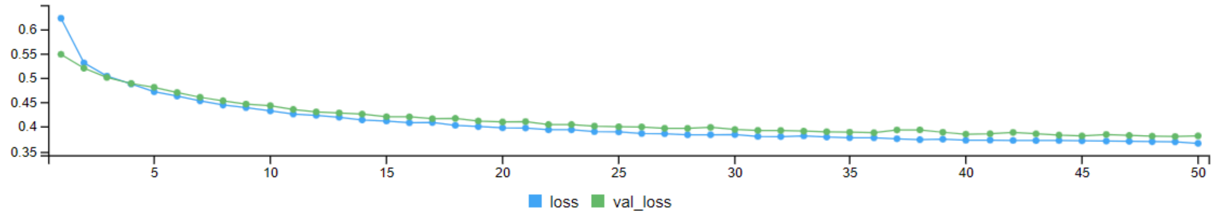
	Model 1	Model 2
Accuracy	0.8438	0.8241
Loss	0.3766	0.4479
AUC	0.9045	0.8862
Calibration	Overfit	Slightly Overfit
Layers	3	4
Units	100, 50, 1	100, 50, 25, 1

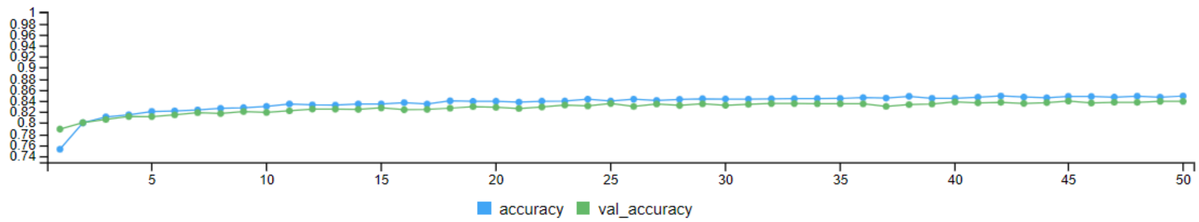
Recommendations:

Model 1 outperformed Model 2 in loss/accuracy and AUC score. However, Model 2 achieved a better calibration curve which is slightly overfitted whereas Model 1 is extremely overfit. I believe the appropriate model for ABC Hotels is Model 2. It only slightly underperformed on loss/accuracy and AUC compared to Model 1 but had a superior calibration curve. We want a model like Model 2 because overfitting means the model is flexible enough for the problem at hand. We can then use methods like early stopping, dropout, and regularizations to make the model more accurate.

Appendix:

Model 1:





Model 2: