**Final Report**

Cameron Gallien

DSE6211OM\_FA2023F1\_Machine Learning & Artificial Intelligence

Merrimack University

Prof. Peter Salemi

October 18, 2023

**Executive Summary:**

This paper outlines a project focused on the development of a highly accurate predictive model for hotel booking cancellations at ABC Hotel locations. ABC Company's objective was to create a model that can precisely anticipate booking cancellations, a crucial factor in optimizing their efficiency and customer satisfaction.

To achieve this, a dataset of over 35,000 hotel bookings provided by the ABC Hotel Company was utilized. These bookings were categorized based on their cancellation status, which served as the target variable for model training. This data was first split into training and testing sets. From there, data preprocessing steps were taken, including feature selection, transforming variables, one-hot encoding, feature scaling, and ultimately, transforming the data into tensors.

Two different dense neural network architectures were explored, with different combinations of hidden layers and units. Additionally, these models were subjected to optimization using various activation functions, optimizers, and loss functions. The first model, featuring two hidden layers with 50 and 20 units, showed promise, achieving an accuracy of 85.57% on the test dataset. Learning curves suggested that this model's fitting to the data was effective, with minimal overfitting. A second model was also created with three hidden layers of 100, 70, and 30 units respectively, and included different activation functions and optimizers. This model produced the highest accuracy at 86.9% on the test data, exceeding the original model's performance. It also displayed improvements in false positive and true positive rates.

Considering the business need, the second neural network emerged as the most suitable choice for ABC Hotels. Its high accuracy and AUC values of 86.9% and .931 respectively, make it a reliable tool for predicting cancellations within hotels. However, with the highest accuracy of either model being 86.2%, it is suggested that further efforts be taken towards developing an even more accurate model.

**Approach & Data:**

This project was approached with one goal; to create the most accurate model possible for predicting cancellations at ABC Hotel locations. To aid in creating this model, the company provided a data set containing over 35,000 bookings for which it was known whether the booking was cancelled. This booking status was used as our target variable and would be used to train the models on what types of bookings have high cancellation rates. By utilizing this information, we were able to train our models to see which factors contribute the most to future cancellations so that the hotel company can be fully prepared for any situations they might find themselves in.

The data was split into training and testing sets so that we would have a separate set of reviews that the models have not been trained on. This test set would be used to simulate how the model would perform on future hotel reservations that the hotel may receive. From there, the models were analyzed to determine which model produced the most accurate results.

Before any models were created, we needed to start with the basic data processing steps that were laid out in the analytic plan and preliminary results reports. The first step was to select which variables should be included in our analysis, and which variables should be excluded. The first excluded variable was the Booking ID as it is an arbitrarily assigned identifier that does not have any specific value for the bookings. The date variable was also excluded from the dataset, but not before creating three additional variables, which were the month and season of the booking, as well as the day of the week that the booking began. Since there were hundreds of different dates ranging from 7/1/2017 to 12/31/2018, this had the potential to introduce a high level of dimensionality that could cause overfitting. By using more simplified date variables, the models should be able to pick up on more general trends such as the time of year, month, or week that they were booked.

In addition, there were multiple categorical variables that needed to be converted to factors so that the model could analyze them properly. These variables are Type of Meal Plan, Room Type Reserved, Market Segment Type, as well as the Arrival Season and Arrival Day of the Week variables that were created in the previous step.

The next step was to process the data so that it was in the format required for the model to properly analyze it. This included splitting the data into training and test sets, one-hot encoding the features, scaling the numerical features so that no one variable outweighs the others, before finally converting the data into tensors. These steps allowed for the models to properly digest the data and produce the highest accuracy levels possible.

Lastly, it was time to decide which models to utilize. The two-layer dense neural network from the Preliminary Results report was included, as well as another dense neural network with three hidden layers. The architecture of these models will be expanded upon in the next section.

**Findings & Model Evaluation:**

The first model utilized was a dense feed-forward neural network with two hidden layers. These two hidden layers contained 50 and 20 units respectively, and both utilized the Rectified Linear Unit (ReLU) activation function. The number of layers and units included in the model were chosen after experimentation with additional numbers of layers and units and comparing the accuracy and loss values that accompanied them. The activation function was chosen because of its popularity as an activation function for deep learning (Chollet et al., 2022, p. 109). The final output layer utilized the sigmoid activation function which allows the model to output a probability (p.109).

From there the model was compiled using the RMSProp optimization algorithm and the binary crossentropy loss function. RMSProp is considered a valuable optimizer for the majority of dense-neural networks, and the binary crossentropy loss function is the most suitable function for binary classification problems that involve probabilities (Chollet et al., 2022, p. 109-110). The final element of the compilation step was choosing accuracy as the observed metric as we would like to see the proportion of predictions that the model accurately identified.

From there the model was trained on the dataset using the training features and training labels that were created during the data pre-processing steps. 50 epochs were used with a batch size of 512, and a validation split of .15 due to the size of the data that was provided. In the preliminary analysis portion of the project we utilized a .25 validation split, but after some experimentation it was found that .15 produced greater results.

In the Preliminary Results report this model was able to achieve an 84.68% test set accuracy, however with the added variable for the booking’s day of the week which was not included in that report, this model improved to an accuracy of 85.57%. This was a promising figure that shows that our model is on the right track to produce highly accurate results for the ABC Hotels company.

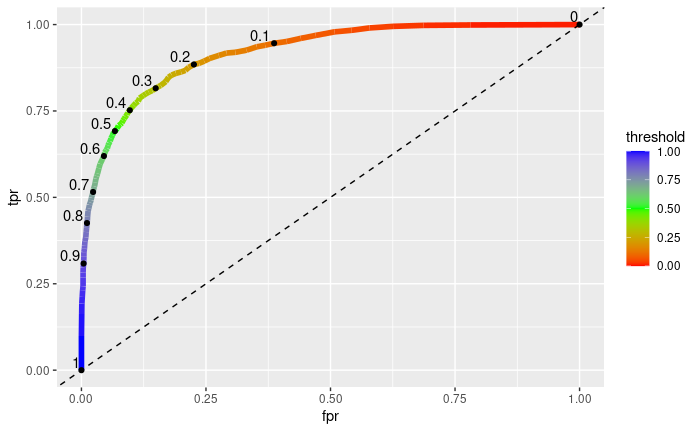
Learning curves were also utilized on both the training and validation sets to verify whether the model was overfitting or underfitting the data.

A graph of a number of objects

Description automatically generated with medium confidence

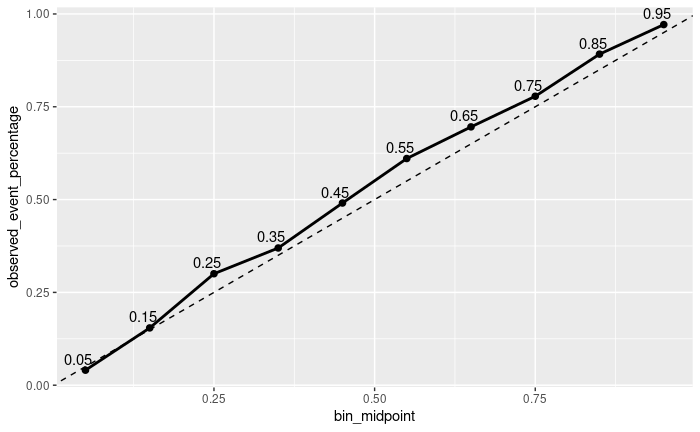
The learning curve also showed extremely promising results. Both the training and validation loss curves converged to a stable low value, and both accuracy curves continuously grew higher which suggests that the model is fitting the data well. In addition, the closeness of the two lines to each other shows that the model is not exceptionally overfitting or underfitting, as they both increase/decrease at a similar rate. There may be a slight amount of overfitting as the training loss is lower than the validation loss, however the overfitting does not appear to be severe. Also the validation loss/accuracy metrics are both at acceptable levels for a useful model.

The model was also evaluated using an ROC curve which was not performed in our original analysis.



The area under the ROC curve (AUC) had a value of .919, which proves that it performs much better than random guessing.

In addition, a calibration curve was utilized to see how well the predicted probabilities were calibrated.



Since the calibration curve lies very close to the dashed diagonal line, we can verify that our original model is well calibrated. Since the curve generally lies above the dashed line, this means that the model is making slightly under-confident predictions in regard to cancellations. This is not necessarily a bad thing as far as the goals of the model. Over-confident probabilities could put the hotel in a bind if there are not as many cancellations as expected and they’ve already taken measures such as cutting staff or overbooking to combat it. While the preliminary results model did a great job of predicting cancellations, it was time to test another model to see if we could improve upon the results.

For this model I decided to try a different activation function and optimizer. For an activation function I chose to use the Scaled Exponential Linear Unit (SELU) instead of the ReLU function used previously. This function was chosen as using ReLU often leads to a problem called “vanishing gradients” which does not occur when using the SELU function (Ahmed, 2022). As for the optimizer, I decided to try the ADAM optimizer, as according to Oppermann (2021) ADAM takes “the best of both worlds” between the RMSProp and AdaGrad optimizers. The sigmoid activation function and binary crossentropy loss function were both still used in the second model.

To start, I wanted to try to create a model that overfits the training data before scaling back the amount of layers, units, and epochs to find a happy medium. The first model attempted consisted of four hidden layers, with 500, 200, 100, and 50 units respectively, and was iterated over 200 epochs. I kept the .15 validation split from the previous model due to the high number of observations provided by the company.

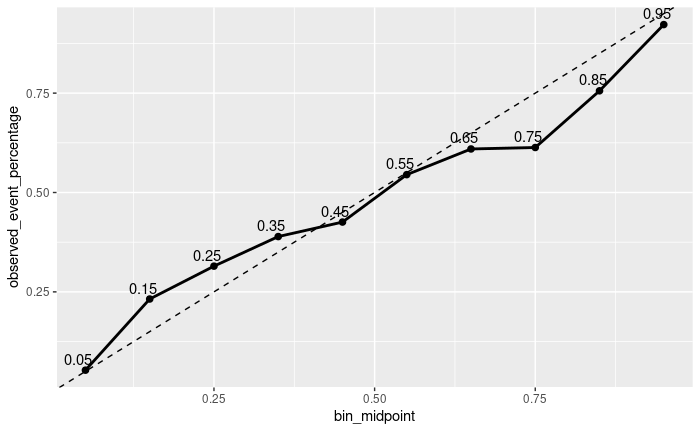
This model definitely suffered from overfitting, as the learning curve below shows:

A graph of loss and loss

Description automatically generated

The training and validation loss curves separate more and more as the number of epochs grows, and the training accuracy is significantly higher than that of the validation accuracy.

In addition, the calibration curve shows additional problems:



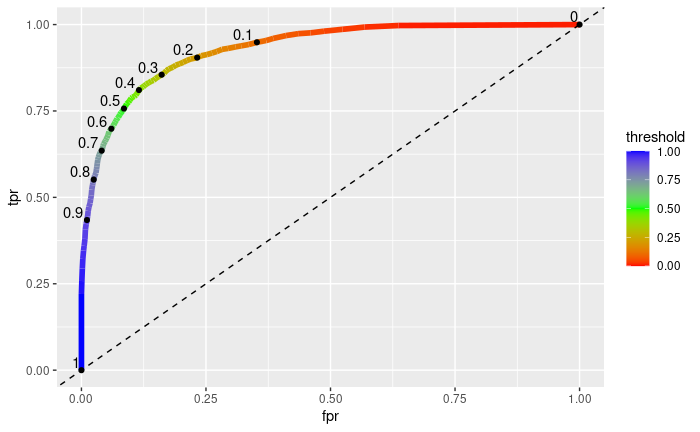
This curve shows that within the interval of 0-.4 (bookings deemed not likely to be cancelled) the model is not assigning high enough probabilities, and within the interval of .4-1 (booking deemed more likely to be cancelled) the model is assigning higher probabilities than is deserved. This model could prove very difficult for the organization to utilize, as it over-predicts cancellations when they are more likely, and underpredicts cancellations when they are not. With that said, this model did produce a slightly higher total accuracy than that of the previous model, with an 85.59% accuracy rate. This was a sign that the new model was moving in the right direction.

After much tinkering with the number of layers, units, and epochs, the final model was created. The final model consisted of only three hidden layers, with 100, 70, and 30 units respectively, and was iterated over 100 epochs.

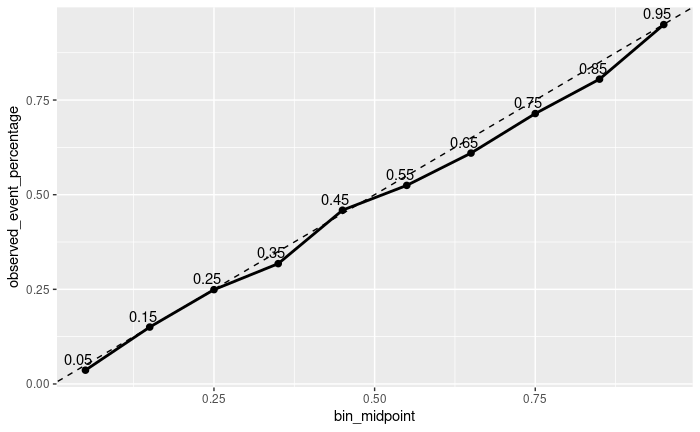
A graph of loss and loss

Description automatically generated

The learning curve above shows that there was a slight amount of overfitting with our data, as the training and validation loss curves do slightly separate as the model iterates over more epochs. On the other hand, the training and validation accuracy curves stay at relatively the same level, showing that the model is generalizing well to unseen data.



In addition, the AUC value for this model was .931, which is also an improvement over the original model. This shows that the model has an even greater ability than the original model to predict cancellations compared to random guessing.



The calibration curve does fall slightly below the dotted line in our plot as the observed event percentage rises, which illustrates a slight overconfidence in predicting cancellations. With that said, the curve does follow the line quite closely, which shows that our model is well calibrated.

In the end, this model produced an overall accuracy of 86.9% on the test data, the highest accuracy of any model that was created. It also showed improvement in regards to the false positive and true positive rates. The original model had an FPR of .083, and a TPR of .729, whereas this new model had an FPR of .082 and a TPR of .768. This proves that our second model is slightly more accurate than the original model that was produced in the preliminary analysis.

**Recommendations:**

After reviewing both created models, it is evident by the results of the model testing that the second neural network produced the greatest overall results. With a test accuracy of 86.9%, and an AUC of .931, there is enough evidence to show that this is the most successful algorithm between the two. In addition, this model showed higher true positive rates and lower false positive rates than the previous model, proving that it is a better fit to be used within the ABC Hotels company.

This model can be leveraged in many ways to improve the operations of the company when there are many reservations with a high probability of cancellation. The hotel can optimize staffing levels to ensure they have adequate staff on hand during peak periods while reducing staff if many cancellations are expected. In addition, the hotel can overbook rooms or offer discounts to increase customer retention during high-cancellation periods. All of these steps can ensure that ABC Hotels are fully prepared for any scenario that may arise.

However, with the best model having a test accuracy rate of only 86.9%, it is recommended that efforts be directed towards refining the model even further. Access to larger datasets, or ongoing model refinement could potentially yield more accurate cancellation predictions, reinforcing hotel management’s ability to predict a high number of cancellations. While the second neural network model does provide a suitable solution for ABC Hotel’s needs, there may still be room for improvement to ensure that all of their hotels are fully prepared for any cancellations that may come their way.

**References**

Ahmed, G. S. (2022, January 14). *SELU (Scaled Exponential Linear Unit) Activation Function*. OpenGenus.org. https://iq.opengenus.org/scaled-exponential-linear-unit/

Chollet, F., Kalinowski, T., & Allaire, J. J. (2022). *Deep Learning with R* (2nd ed.). Manning.

Oppermann, A. (2021, December 21). *Optimization in deep learning: AdaGrad, RMSProp, ADAM*. artemoppermann.com. https://artemoppermann.com/optimization-in-deep-learning-adagrad-rmsprop-adam/