**Preliminary Results**

Cameron Gallien

DSE6211OM\_FA2023F1\_Machine Learning & Artificial Intelligence

Merrimack University

Prof. Peter Salemi

September 30, 2023

**Introduction**

The ABC Hotel company is looking for a way to identify bookings that carry a high risk of cancellation. This will allow the hotel to prepare for cancellations when bookings that meet the criteria are identified, and potentially target these customers with additional advertisements and offers in an attempt to prevent the rooms from being cancelled. This paper presents a supervised classification process to provide this business need, leveraging a data set containing over 35,000 past bookings for which it is known whether or not the customer’s booking was cancelled.

The target variable utilized for this supervised classification problem is the booking status of the reservation. Each reservation has been marked as “cancelled” or “not cancelled,” which was used to train the model as to which reservations are more likely to be cancelled. By utilizing this information we will be able to input the data into our algorithm and train it 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.

**Data Processing**

The first step in processing the data 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 two additional variables, which were the month of the booking, and the season of the booking. 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. Also, the specific dates would correspond to different days of the week than they would in future use, so they could make the model less accurate if there was too much of a focus on those specific dates. By simplifying the date variable the model should be able to pick up on more general trends such as the time of year or month 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, and the Arrival Season variable that was 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 model to properly digest the data and attempt to produce the most accurate model possible.

**Model Training and Compilation**

The architecture chosen for our model was a dense feed-froward neural network with two hidden layers. These two hidden layers contained 50 and 20 units respectively, and both utilized the 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 .25 due to the size of the data that was provided.

**Model Evaluation**

The first step in evaluating the model is the most obvious, test set accuracy. The model was able to achieve an 84.68% test set accuracy, which is pretty encouraging for a preliminary model. While additional steps can be taken to achieve higher accuracy, this was a promising figure that shows that the 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 data

Description automatically generated with medium confidence

The learning curve also showed extremely promising results. Both the training and validation loss curves converge to a stable low value, and both accuracy curves continuously grow 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.

**Proposed Next Steps**

Based on our evaluation of the preliminary model, there are multiple steps that can be taken to further refine it. First and foremost, fine-tuning the hyperparameters used in the model may provide improved results. This includes fine-tuning the number of hidden layers, units per layer, number of epochs, and percentages of training/test data among other things. Finding the most precise value for these hyperparameters could vastly improve the accuracy and loss rates shown in the learning curve.

In addition, other activation functions and optimization algorithms could be experimented with as they may provide greater accuracy on our original model. Alternative activation functions like selu could provide better results, especially if additional layers are added. There are also additional optimizers that could be utilized, such as SGD & Adam, which may show additional benefits when implemented into the model.

**Conclusion**

Once implemented, this dense feed-froward neural network will address the stated business needs of ABC Hotels in regards to predicting future cancellations among their hotels. The preliminary model showed promising results but requires further refinement and tuning before it can be fully implemented for use within the company. The proposed next steps of tuning the model’s hyperparameters and altering the activation functions and loss algorithms aim to improve the model’s accuracy, allowing more precise predictions regarding the chances of a hotel reservation being cancelled.

Ultimately, the goal of the project is to provide hotel management with a reliable tool so that they can properly adjust the business needs of their locations. This will enable them to make data-driven decisions, optimize their booking management strategies, and enhance customer satisfaction by minimizing cancellations. This will allow ABC Hotels to improve revenue and customer retention while ensuring a seamless and enjoyable booking experience for their guests.

**References**

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