Neural Network Models for Customer Spending Patterns

Richard Feng
rzfeng@wisc.edu

Abstract

We compare deep learning and machine learning learning methods for a classification problem on a large tabular dataset looking to predict what spending category a consumer will be in based on various consumer attributes, including age group, gender, occupation group, and more. It was found that a basic Softmax Regression architecture as well as the Multilayer Perception (both using Cross Entropy loss functions) could not outperform a random forest classifier, largely due to our underlying dataset's explanatory variables not having much variation. The random forest classifier largely only used one explanatory variable, Product Category 1, in making its classification, as this was the only variable in our dataset that had true variation. We experimented with different number of spending categories for our prediction, as well as tuning hyper-parameters for each architecture, and the Random Forest Classifier still achieved a better testing accuracy than our neural networks. This was especially true when we used 2 and 10 buckets, as random forest achieved 81% and 37% test accuracy, respectively, whereas our Softmax Regression architecture (the better of our two architectures) had only 63% and 37% testing accuracy. The basic MLP architecture consistently performed the worst out of the three.

1. Introduction

Last year, the average American shopper spent nearly $1,000 dollars for a total of 717.5 billion dollars on Black Friday which represents nearly a 4.5 percent increase over the previous year. In order for companies to try to maintain and capture this growth for increased profits, companies need to be able to predict how much a shopper would spend during Black Friday. This is difficult because consumers are influenced by a number of factors. However, if companies are able to do so, they can adjust their marketing campaigns, change prices, and offer more compelling discounts to optimize prices in order to increase their bottom line.

This paper uses three different deep and machine learning techniques to try to classify consumers Black Friday spending in an unknown store into distinct bins using features such as age, marital status, gender, and others. After cleaning and normalizing our data, we experimented with Multilayer Perceptrons, Softmax Regression, and Random forests in order to discern the most precise solutions. Both neural network architectures used a cross-entropy loss function, as this type of loss constantly outperforms MSE for classification problems.

Our results were varied. Even after tuning our models and testing different amounts of spending buckets, our deep learning models, Multilayer Perceptron and Softmax Regression, were relatively ineffective in generating correct predictions. On the other hand, our machine learning Random Forest Classifier performed admirably and had higher accuracy for classifications for each number of spending buckets.

2. Related Work

The want to use deep and machine learning techniques to classify consumers monetary spending tendencies have not been new. Previous neural network work includes the Multilayer Perceptron used to study banking services by Alborzi et al.[1] and the Multilayer Perceptron used to study online shopping by Crone et al.[3]. Toth et al.[4] used the deep learning technique of recurrent neural network to predict shopping behavior. Previous machine learning work includes the random forest clickstream classifier used by Awalker et al.[2] to predict consumer preferences. After reading through these and other papers on Softmax Regression, we decided that the most efficient way to solve our problem was to evaluate a MLP model, Softmax Regression model, and a Random Forest Classifier and work off the best performing model.

Most of our approach is based off of these previous papers. Although Alborzi et al.[1] and Crone et al.[3] both use Multilayer Perceptrons, we applied these techniques to a different type of classifications relating to consumer purchase volume and behavior. This paper also does a comparison of different techniques rather than just optimizing a single one.

3. Proposed Method

For all of our models, we chose to use cross entropy as our loss function. Cross entropy loss (formula shown below) is known to work well for classification problems, and incorporates maximum likelihood estimations. In the 1980s and 1990s MSE was a popular loss function. However, it has been found that using cross-entropy loss results in faster and more robust learning for the model, especially for sigmoid and Softmax Regression activation functions.

We firstly implemented a Multilayer Perceptron with one hidden layer containing 15 neurons, using different activiation functions of tanH, sigmoid, and reLU. We ended up using tanH because it converges quicker than sigmoid while ReLu resulted in a vanishing gradient.

Afterwards we used a Softmax Regression model; a simple two-layer neural network with just one input and one output layer. Here, our activation function is Softmax Regression and the model uses a multi-class cross entropy loss function.

Our Random Forest Classifier implementation stems from scikit-learn, a library that provides many machine learning tools.

Using each of these models, we train it to predict 2, 4, and 10 different classes on customer spending amount and evaluate the models and results.

4. Experiments

4.1. Data Cleaning and Preprocessing

The dataset we used came from Kaggle, titled "Black Friday" and posted by Mehdi Dagdoug. The data contained 550,000 observations of shoppers from a single store on Black Friday. Because of the nature of the data, the actual store was not given in order to remain anonymous, likely to protect the privacy of the customers. The dataset contained a bunch of categorical variables relating to the customer: gender, age group (with the buckets seeming to be rather arbitrary), occupation group (1-20), marital status; as well as data regarding their purchase, such as product category (1,2, and 3) and the amount purchased. Once again because of the sensitivity of the data, most of these variables were encoded as integers and the dataset did not provide a key to each of these values, so it was tricky to explore the dataset and know exactly what we were working with.

In order to prepare the data for analysis, we first had to clean the data, which involved removing the second and third product categories because they contained large amounts of missing values. In addition, we converted price into bins as we wanted to create a classification problem (we tested our networks with different amounts of pricing bins). In addition, we one-hot encoded each categorical data type, leaving us with binary variables for each categorical feature, and finally it was necessary to do some data pre-processing to convert the data to the correct PyTorch tensor data types.

The purchase variable (that we are trying to predict) was relatively uni-modal around 5000-8000. Most of the explanatory variables in the dataset were not useful. However it is slightly skewed and has clusters of purchases in the 15000 and 20000 range.

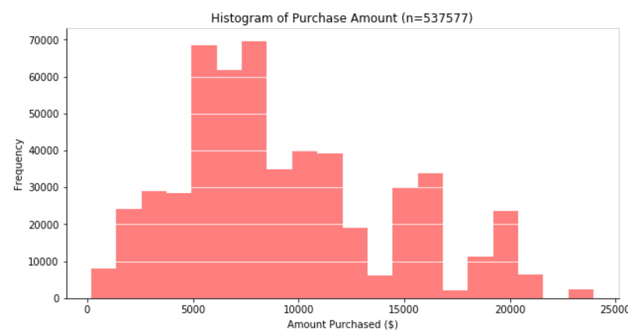

Figure 1.

Most of the categorical data was largely unhelpful, as they contained very little variation with amount spent. For example, both genders and each age group on average spent nearly the same amount. However, the one insightful feature we had was Product Category 1, where many of the different product categories had starkly different averages for amount spent. It appears as if this is why we see some clusters in the distribution for amount purchased.

In addition for our data pre-processesing, we normalized each feature within the dataset. This should help our network converge toward minima quicker during gradient descent.

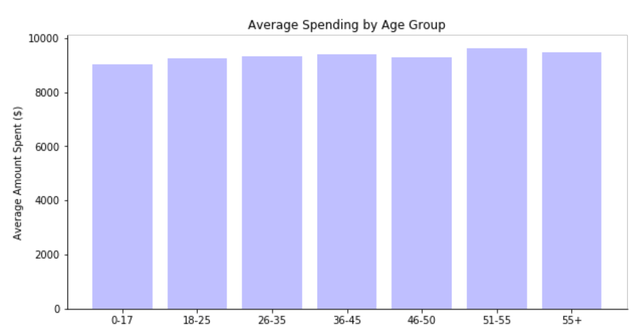

Figure 2.

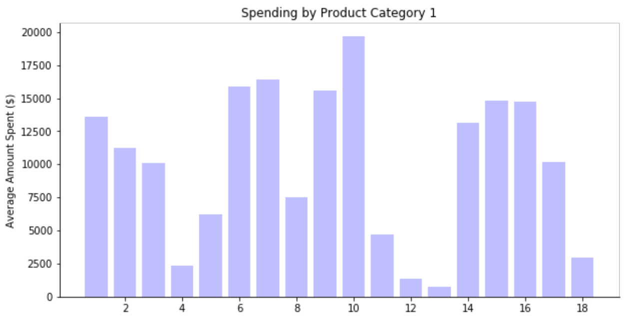

Figure 3.

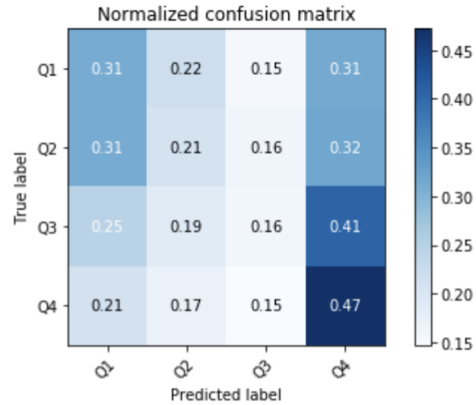

Figure 4. Confusion matrix of the MLP 4 class predictions

4.2. Software

All three of our models were written in Python. For our MLP and Softmax Regression, we used Pytorch, a deep learning library. Pytorch heavily aids us in mathematical operations needed for our models, and provides many other useful functions for making deep learning models from scratch much easier. For our Random Forest Classifier, we used the library scikit-learn to create and train our model, as well as to create the confusion matrices. For our exploratory data analysis and data preprocessing, we used the libraries Pandas and Numpy.

4.3. Hardware

To speed up the training time on our two deep learning models (MLP and Softmax Regression), we utilized cloud computing resources. In the past, we have found that Kaggle's GPU trains deep learning models faster. Kaggle offers access to Nvidia Tesla K80 GPU. Interestingly, Google Colab's free GPU also offers this same GPU, even though it has been running slower than Kaggle. However, Kaggle frequently crashes without saving, so frequent saving and saving backup copies locally was necessary.

5. Results and Discussion

Three final models were evaluated over the course of this study, including a Multilayer Perceptron, a Softmax Regression model, and a Random Forest Classifier.

Using our Multilayer Perceptron with a tanH activation function, the testing accuracy was 28.73% with 4 possible classes to predict. In the confusion matrix below, we can see how the model does relatively well predicting class 1 and class 4, the two extremes. However the two intermediate classes, Q2 and Q3, were predicted almost randomly, which raises some concern in how our model is learning.

Q3 seems to be the hardest to predict, as only 16% of Q3 customer labels were predicted as Q3, while the majority of the predicted labels actually were Q4. Q2 was also hard to predict, since 21% of class 2 labels were predicted as class 2, but 31% of them were predicted to be class 2 and 32% were predicted to be class 4.

This model is clearly very unreliable in predicting Q2 and Q3. If we were to use this model, then the best use would probably be to predict is a user is going to spend a massive amount of money (i.e. a customer in Q4). This leads us to implementing a binary classification model later on.

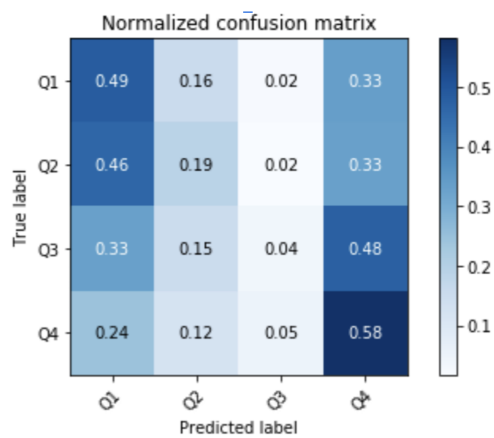

Figure 5. Confusion matrix of the Softmax Regression 4 class predictions

For our Softmax Regression model, we were able to increase our testing accuracy to 56%. The Q1 predictions are a lot more accurate as well as the Q4 predictions by almost 20% and 10% respectively. Although it's a 27% jump in accuracy compared to our MLP, we can see from the confusion matrix below that the predicted labels are still not very good at predicting the correct classes for Q2 and Q3.

However, even though only 19% of Q2 labels were predicted correctly, the the majority of the other predictions were in Q1 rather than Q4, indicating that this model may perform better when doing binary classification later on. Similarly, Q3 only predicted 4% of the Q3 labels correctly, but the majority of the predictions were Q4, rather than Q1, also indicating that this model may be better than our MLP model at classifying the two extremes.

With our Random Forest Classifier, the diagonals in the matrix are much more distinct, allowing us to see at a quick glance that this model performs the best so far. This confusion matrix in general also shows the best results, because if a customer spent a large amount of money (i.e. Q4), the model would rarely predict Q1 and Q2. In the previous confusion matrices, we can see that even though a customer was in Q4, the model would still frequently predict Q1 which is very far off.

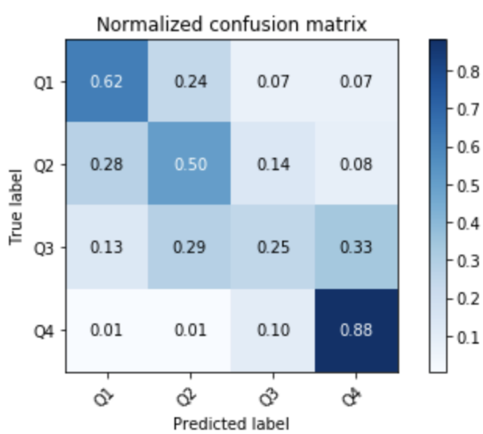

Figure 6. Confusion matrix of the Random Forest Classifier predictions on 4 classes

We follow up by converting our Problem into a binary classification problem by adjusting our models to have 2 output classes. This model would be able to tell us whether a customer would be spending a large amount of money. or a smaller amount of money

Starting with our MLP with a tanH activation function, our test accuracy was about 59%. When observing the corresponding confusion matrix, we can see our model does a little better predicting customers who will spend a lot than customers who will spend less.

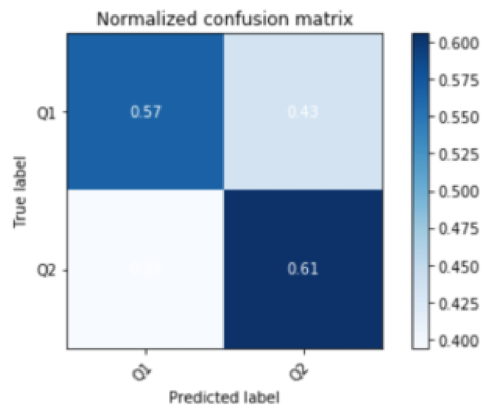

Figure 7. Confusion matrix of the MLP predictions on 2 classes

When we move to our Softmax Regression model, we can see our testing accuracy is almost 63%, and looking at the confusion matrix this model was able to predict lower spending customers relatively well.

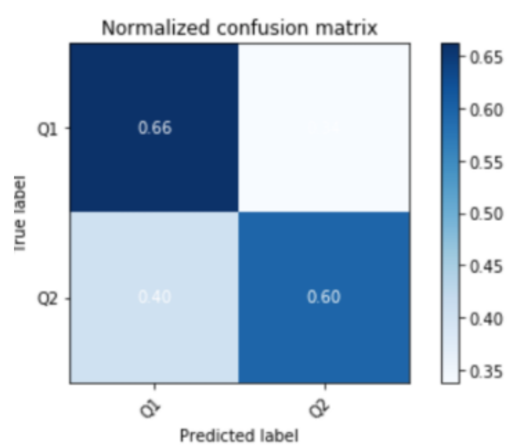

Figure 8. Confusion matrix of the Softmax Regression model predictions on 2 classes

Finally with the Random Forest Classifier, we get a nice testing accuracy of 81%. Observing the confusion matrix, this model (similarly to our Softmax Regression model) also predicted customers in the lower spending bracket better than higher spending customers.

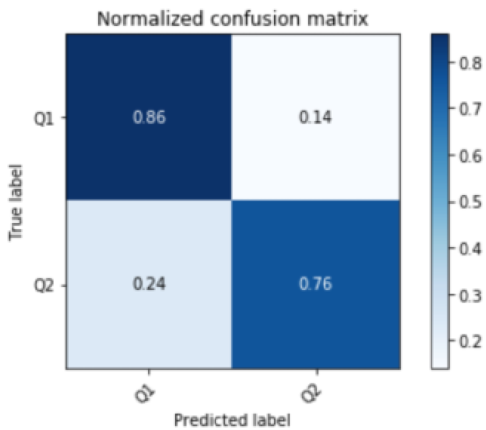

Figure 9. Confusion matrix of the Random Forest Classifier predictions

For further inference, we decided to see how well each of the three models performed using 10 classes rather than 2 and 4 classes. The results solidified how our Random Forest Classifier is indeed the best model, as the diagonal in the confusion matrix was significantly more bold than the other two models. We also saw how MLP and Softmax Regression had similar results to predicting 4 classes, where the extreme classes Q1 and Q10 were predicted more frequently than the in between classes Q2-Q9.

To try and reason why our Random Forest Classifier performed the best, we examined the importance that each feature contributed towards our model. The most important feature was found to be Product Category, followed by Occupation and Marital Status. However, Produce Category was clearly the number one biggest contributor to the Random Forest model compared to occupation and marital status. Although we are unsure of what Product_Category_1 is exactly, we can speculate that it may be the product category of an item or items a customer purchased. This is then logical that using since some product categories may have higher prices compared to others. After training our model without using this feature, our Random Forest model's accuracy dropped significantly by about 20%.

	importance
Product_Category_1	0.863938
Occupation	0.080526
Marital_Status	0.009674
City_Category_C	0.004236
Stay_In_Current_City_Years_1	0.003787
Stay_In_Current_City_Years_3	0.003391
Stay_In_Current_City_Years_0	0.003150
Stay_In_Current_City_Years_2	0.003082
Stay_In_Current_City_Years_4+	0.002878
Age_36-45	0.002737
Age_26-35	0.002736
City_Category_A	0.002691
Age_18-25	0.002637
Gender_M	0.002517
Gender_F	0.002378
Age_46-50	0.002185
Age_55+	0.002107
City_Category_B	0.002066
Age_51-55	0.001972
Age_0-17	0.001312

Figure 10. Importance of each feature from our Random Forest Classifier

6. Conclusions

Given our dataset with very vague and general features such as gender, age bracket, martial status, and others, our deep learning models were unable to accurately identify patterns to predict customer spending patterns.

The results of all our models predicting all 3 different amount of classes are shown in the table below.

	Test Accuracy (2 classes)	Test Accuracy (4 classes)	Test Accuracy (10 classes)
MLP	~59%	~29%	~13%
Softmax Regression	~63%	~56%	~16%
Random Forest	~81%	~56%	~37%

Figure 11. Table of accuracy results

Our Random Forest Classifier was our best performing model. This may be due to the inherent feature extraction methods this model has. It was able to identify the most important features and use it to its advantage in producing the most accurate results. Our deep learning models may not have yet caught on to the most important features in terms of predicting consumer behavior.

However, seeing as occupation, martial status, and city category were the next important features in our best model, these may be some key sources of information one may want to extract if they want to predict customer spending habits. Whether or not they were a local (i.e. stay in current years) may also have an effect on consumer spending patterns. If a person has lived in the area for a while and is comfortable there, they may be willing to spend more money with their sense of security. A person who is new to the area may not be as comfortable, so they aren't willing to spend as much.

Similar to previous papers, we analyzed using different architectures and models to approach a problem before attempting to optimize one of them. It may take a a while to tweak the MLP model to get a test accuracy of 60% for 4 classes, but after comparing the results, working on the Softmax Regression architecture would deem quicker and more efficient in achieving higher accuracy. In our results, since the training and testing data were similar, we suspected no overfitting was happening, so variations such as using dropout were not used. The next step to work towards this would be to add more layer to our Softmax Regression model and use dropout or regularization if necessary to achieve a good model.

In the future, one thing we would like to account for is the fact that the bucket for amount purchased are ordinal, which our networks and classifiers did not take into account. Because we did not treat our data as ordinal, the difference between classifying an observation in bucket 1 (the lowest spender) that should belong to bucket 4 (the highest spender) and classifying an observation in bucket 1 and bucket 2 would be treated the same, which does not make since in practice. The loss should be greater for predictions that were further off, thus in the future we would want to treat this as an ordinal linear regression.

References

[1] M. Alborzi and M. Khanbabaei. Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed rfm analysis method. *International Journal of Business Information Systems*, 23(1):1–22, 2016.

[2] A. Awalker, I. Ahmed, and T. Nevrekar. Prediction of user purchases using clickstream data. *International Journal of Engineering Science and Computing*, 2016.

[3] S. F. Crone and D. Soopramanien. Predicting customer online shopping adoption-an evaluation of data mining and market modelling approaches. In *DMIN*, pages 215–221, 2005.

[4] A. Toth, L. Tan, G. Di Fabbrizio, and A. Datta. Predicting shopping behavior with mixture of rnns. In *ACM SIGIR Forum. ACM*, 2017.