HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Applications of Machine Learning to Probability of Students Continuing to College

Nathan Chen

Archbishop Mitty High School, USA

ABSTRACT

We attempt to use machine learning to predict whether or not a student will end up going to college. The data that is provided are statistics such as whether their parents went to college, whether they live in a rural or urban area, and the students' interest level of actually going to college. The main object of this project was to apply machine learning to accurately predict whether a student would continue to college. How this can be useful in the world is that based on certain statistics, a student can see if they are more likely to pursue college or if they are more likely to not continue to college. The conclusions were that the model had been trained to predict pretty accurately.

Introduction

Machine learning is the use of computer technology in order to have the computer adapt and learn datasets without extra code so that it can hopefully make successful predictions.

The dataset used for this project contained information about the students' personal life. The features that were included for this were the type of school, school accreditation, gender, residence, parent salary, average grades, whether or not their parents went to college, and the students interest levels on going to college. The data that is provided is mixed as some are categorical such as the type of school, school accreditation and gender and can not be a numerical value whereas other parts of the data set like grades and salary are numerical. The label for this project was whether or not the student ended up going into college; this left two options one being they went to college and the other being that they did not, this makes this a binary classification. The following is the original dataset before it was cleaned.

type_school	school_accreditation	gender	interest	residence	parent_age	parent_salary	house_area	average_grades	parent_was_in_college	will_go_to_college
Academic	A	Male	Less Interested	Urban	56	6950000	83.0	84.09	FALSE	TRUE
Academic	A	Male	Less Interested	Urban	57	4410000	76.8	86.91	FALSE	TRUE
Academic	B	Female	Very Interested	Urban	50	6500000	80.6	87.43	FALSE	TRUE

Vocational	B	Male	Very Interested	Rural	49	6600000	78.2	82.12	TRUE	TRUE

Both label and feature are shown in this dataset and that is so when the computer later uses this dataset it can determine what features may have led to the outcome of a certain label. The machine can learn and adapt using this dataset so that in the future it can attempt to make accurate predictions. The cleaning of the data included turning a lot of the categorical data such as gender into 1's and 0's. This happened because of many machine learning algorithms needing numerical values and using a binary matrix. This process also included deleting some features that may not have been objective or clear such as house area and the parent age. Another step to cleaning the data was one hot encoding the interest level; this goes back to the reason for fitting with the algorithm and fitting the binary matrix.

Model Prediction and Log Loss

The problem that this project is covering is a binary classification problem. The label results are either that the student goes to college or they don't, meaning there are only two possible outcomes. Due to the fact that this is a binary classification problem, instead of predicting a continuing value that regresses, the machine will train to predict one of the two options and check its accuracy.

Since this is a binary classification problem it doesn't check how close a value it produced is to the correct value when it is training but rather it checks to see how accurate it was in choosing between the positive and negative labels. Instead of using square error like other machine learning algorithms, binary classifications use something known as log loss. Log loss is technical since it has the machine trained by predicting which of the outcomes it is and then giving feedback on how it did. In simpler terms an analogy that can be used about log loss is having someone guess a mystery object on whether it is a pencil or not. Imagine tracking each of their guesses and attempts, Log Loss is a way to measure how "surprised they are". The surprise level in this case means that if they guessed something closer to reality (the pencil), they will not be as surprised compared to guessing something further from reality, they would be very surprised. Log loss is measured in where the closer the machine is to the right outcome, the log loss will not be high whereas if the machine's prediction is far from the right outcome, the log loss will be high. The entire goal of machine learning is to have the log loss as minimal and minimum as possible to reduce error in the predictions. With log loss being minimized, the prediction should be as close to reality as possible allowing it to be more accurate.

The official Log Loss formula is:

$$\text{Log Loss} = -(y \cdot \log(p) + (1-y) \cdot \log(1-p))$$

Where:

- y is the actual binary label (0 or 1).
- p is the predicted probability of the positive class.

Training and Validation Sets

For the model, I split my dataset into two groups, the training and validation set. My validation set had 25% of the data whereas my training set had the other 75% percent of the data. The training set separated some data so that the model can train and learn from its errors so that it can eventually lower that error. The validation set is tested after the training set to see how the accuracy improved and review how the model does on new data after being trained. If the model performed well on the validation set as well as the training set, it means that the model works efficiently and can reach a high accuracy for whatever it is trying to predict.

Binary Classification

Mentioned earlier, this project type is a binary classification, the reason why is since there are only two outcomes, the binary classification algorithms are more suited and meant for problems like these. Problems that have the label choosing between two categories rather than creating its own number. Contrastingly another common machine learning technique is linear regression. The reason why linear regression wouldn't work is because linear regression is meant for predicting numerical values that range outside the realm of 0 and 1. Binary classification is more suited for this project because it is about probability, it is about what factors that a student has and how it will affect their future on going to college. This project doesn't predict a percentage on the likelihood of the student going to college but rather categorizes it in that the student will or won't go.

Neural Networks

Before discussing the importance of neural networks and the purposes they bring, the understanding of what a neural network is should be discussed. A neural network is made of primarily two layers: input and output. The input layer takes in raw data and the number of nodes that the input layer has is determined by the data. Nodes are the foundation and what makes neural networks as they receive input, process them and give back an output. The other essential layer is the output layer which can be a binary classification output, a regression output or a multi-class classification output and it basically gives back a value after the data from the input layer has been processed.

Neural Networks are an important aspect of Machine learning as neural networks are capable of learning and extracting complex data from the datasets and utilize them towards the training. Neural networks can create connections and find patterns among the dataset. This is extremely useful with a binary classification problem because it can find a pattern that helps fit in with one of the options better and through that, the predictions will be more accurate. Since our problem is non-linear, activation functions such as tanh, sigmoid, and relu are really helpful as they can understand complex boundaries. Another reason why neural networks are good is because they have high adaptability and can learn from a multitude of data sets which makes them versatile for binary classification.

My specific project has an input layer, two hidden layers, and a binary classification layer. To give more information about the specific type of output layer, the binary classification layer uses the sigmoid activation function that shows a value between 0 and 1 to represent the probability of being in one of the two classes. Binary classification which best suits my project type is why the model primarily used sigmoid as its output function. The two hidden layers were relu and tahn and which had 8 and 4 nodes respectively. The hidden layers were added because the model wasn't getting very accurate results meaning that the training or the processing through the neural network wasn't great at the time; adding the hidden layers allowed more nodes to process the data and figure out a pattern which eventually led to more accurate results.

Applications to our Data Set

We applied Log Loss and neural networks to target this binary classification problem as discussed before. This was to analyze based on certain statistics what was the likelihood that the student would go to college. The model was trained to predict whether the student was going to a college or not. The following is the code for the model.

```
model = Sequential()
model.add(Dense(8, activation='relu'))
```

```
model.add(Dense(4, activation='tanh'))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy', metrics = ['accuracy'])
model.fit(X_train_scaled,y_train,epochs=600,verbose=0)
J_list = model.history.history['loss']
plt.plot(J_list)
pd.DataFrame(model.history.history)
```

The binary crossentropy error for the training set was 0.3033568. The accuracy of the training set was 0.87. The model was tested on the validation set after it was trained on the training set. The binary crossentropy error for the validation set was 0.44913447. The accuracy of the validation set was 0.81 which is pretty close to the training set.

Overfitting can be a huge problem in this model. Overfitting in this model would mean that it has learned the training data so intensively that it can not accurately measure new student's data. For example it may capture something in the training set that doesn't represent a trend or pattern going on for students. This can lead to unreliable predictions for students and end up having the model become inaccurate. To overcome overfitting, there has to be the proper number of epochs, to ensure that it performs consistently well on the training and validation sets. For this specific project 600 epochs was a proper amount that led to consistent results within the training and validation set.

Conclusion

The machine learning model, which was developed using TensorFlow and Keras libraries, serves as an informative resource for predicting a students' likelihood of going to college. The model is trained and validated with a total sample of 1000 students who each have different statistics such as their interest level on going to college, what type of school they go to, and etc. The training and validation set were split in 75% and 25% respectively. The model predicted the training and validation set pretty similarly as the training set had an accuracy of 0.87 and the validation set had an accuracy of 0.81. The model did not show any signs of overfitting as the data was cleaned and prepared before using it for the model. With all these factors, the model should predict accurately whether or not a student will go to college or not when it is applied to a new student. Some limitations of machine learning is that the model can only train off the dataset that it is given. This means that there are multitudes of factors that weren't taken into place which could have affected the accuracy or how the model trained. This model can be valuable because if a student is deciding whether they want to go to college or not, this model can help them predict what is the likelihood of them going to college and that can help the student make a decision.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.