Pitch Pine Mortality Classification

Nasir, Abdullah
Rowan University
Glassboro, Glouchster County
nasira39@scarletmail.rutgers.edu

Mahmut Gemici
Rowan University
Glassboro, Glouchster County
gemici34@students.rowan.edu

***Abstract*—Data Science is becoming more and more prominent in forestry, there is a huge potential to use data science tools to make better planning decisions. This paper discusses our findings of Pitch Pine Forest data, and our efforts to develop an algorithm predicting mortality of sampled tress based on pre- and post-fire data. We were able to do the necessary pre-processing of the data and then use a one-layer neural network to make high accuracy predictions. Due to limited dataset, we tried oversampling and under sampling to verify our model, we received similar high accuracy with those data manipulation techniques. We also shared our model with a state Forester and received positive feedback. Our primary take away from this was to show how Deep Learning techniques can be used to make critical planning decisions if sampled datasets were scaled to forests across the state.**

Keywords—Data Science, Pitch Pine, Forestry, Neural Network (*key words****)***

I. Introduction

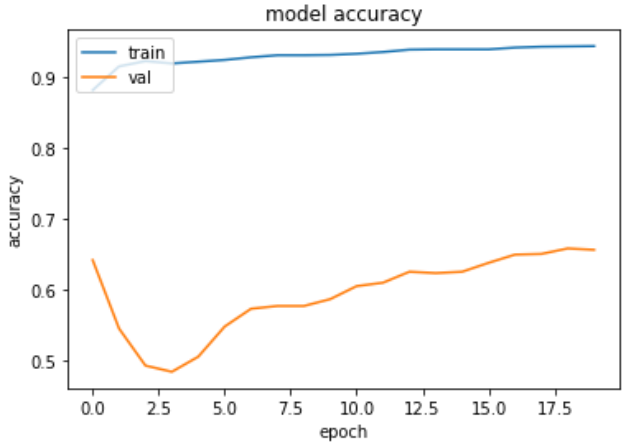
Forest lands make up roughly 42 percent of New Jersey land use, hence requiring major Forestry service efforts to preserve and maintain forestland in Garden State. Computer Science and data science provide tools that are enhancing our capabilities across all fields, our goal with this project was to provide a preview of how Deep Learning techniques can be used in State Forestry. We reached out to a State Forster and acquired a Pitch Pine Forest Data, the dataset primarily consisted of 10 columns with 8 features and one label column. The dataset contained rows of features for each tree with samples collected in 2013 and 2014. 2014 status represents the mortality status of trees post a forest fire and hence our goal was to predict the 2014 status based on features collected in 2013. This can be a very useful tool because forestry actively engages in thinning and controlled fires to prevent massive future fires, so if given a tool that can identify tress at high risk during a forest fire, it can help foresters make better decisions. The Dataset contained less than 2700 rows, so we knew we were limited in that aspect, but deep learning techniques are easily scalable and perform even better with bigger datasets. So, this dataset was a good start, and tabular data is optimal for machine learning techniques and can be used to make highly accurate predictions. We suspected that a one layer or a small neural network could work well with this dataset, as this seemed like a logistics regression problem.

II. Methods

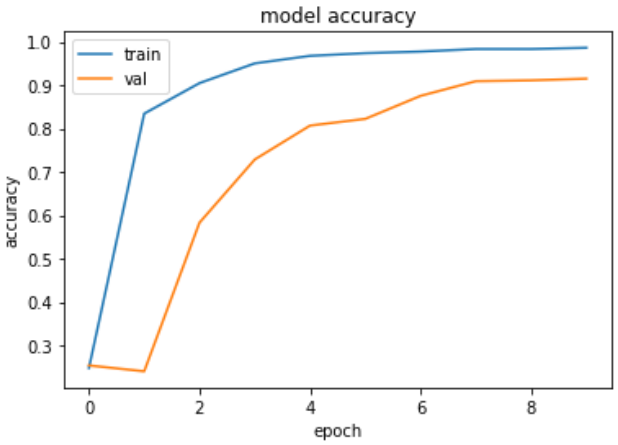
Our first step was to convert string fields into integer fields, dataset contain a few string fields describing status and type of trees, we explored different techniques for this such as Label Encoding and One Hot Encoding, we ended up choosing one hot encoding by using dummies value approach [1]. We also noticed that our dataset was very asymmetric meaning only 13 percent of tress were labeled “dead” and overwhelmingly number of labels were “live” trees. To account for this, we looked at different oversampling and under sampling techniques to normalize the bias, we decided to oversample minority labels using “imbalanced” python library to duplicate rows and equal the amount of live and dead trees [2]. For our next step we decided to use a one-layer neural network with two dense layers, we set the input dimensions of first layer equal to our number of feature columns and one dimension for second layer. We used “relu” activation for first layer and ‘’sigmoid” for second layer since it’s a binary classification problem. We used “adam” optimizer and a “binary cross entropy” loss function from sklearn library. Our hyperparameters that we could tune included validation split size, batch size and number of epochs.

III. Results

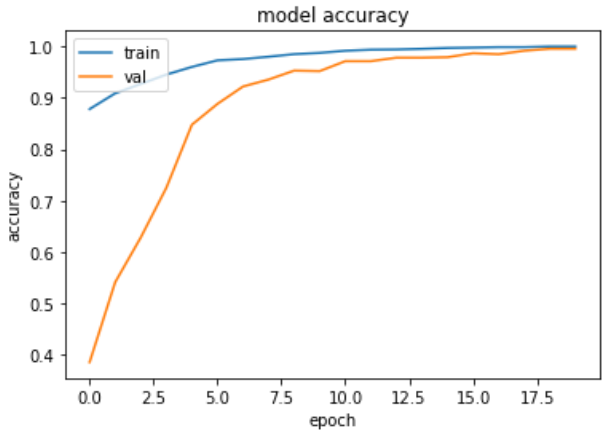
Batch Size had the biggest impact when tunning our hyperparameters, we started with a 256-batch size found that higher batch size led to poor accuracy results despite having high training accuracy, as seen below.



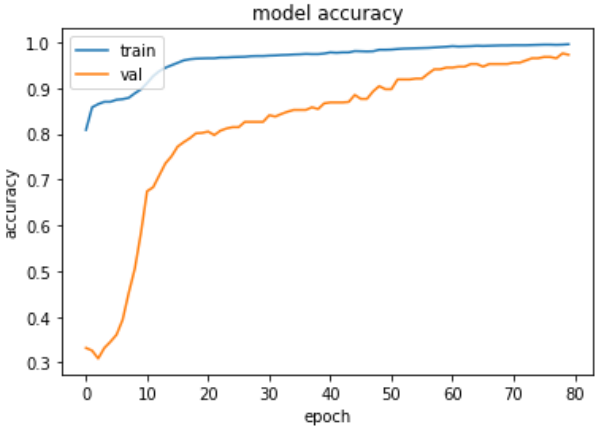
When we used a lowered batch size while keeping other hyperparameters untouched we received much better results as see below. We believe since we have a small dataset using a smaller batch size results in better learning curve for the model.



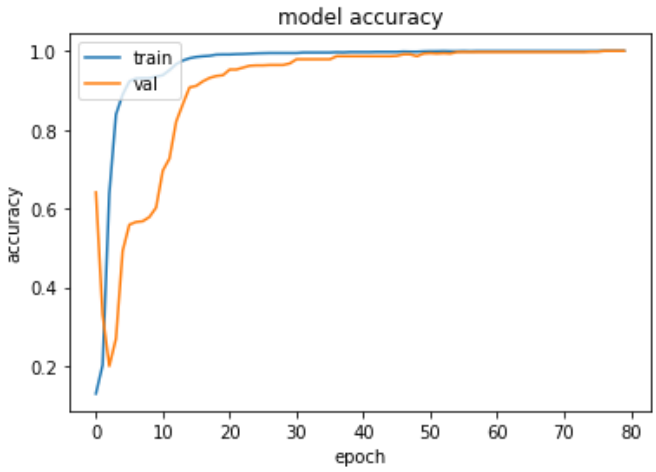
As you can see model still reaches a high accuracy despite having low epochs, now let us compare this to large batch size with higher number of epochs.



Number of Epochs also had a big impact on accuracy of the model, we used a smaller batch size and lower epochs number and compared it with using a large batch size but higher epochs to see if the model accuracy eventually converges. Below you see the results of low epochs but with a very small batch size.



As we can see model does eventually reach higher accuracy numbers, so we conclude that using a smaller batch size with higher epochs would lead to optimal dataset. When we implemented that on our model, we reached an accuracy of 100 percent, but we suspect since we have a small dataset the model starts to overfit, but we believe with a larger dataset these hyperparameter configuration would give the best results.



IV. Discussion

Forestry collects large amount of tabular dataset every year and inventory datasets date back to decades, if that data collection can be tailored for machine learning tools that can result in highly accurate predictions for several problems. Machine Learning methods can be expanded into various number of problems such as calculating Carbon sequestration numbers in forests as well as tree classification using aerial imagery. Our goal with this project was to demonstrate how we can make accurate predictions given sampled data and how there is a potential to scale these efforts across all forestlands. We were able to achieve high accuracy because tabular dataset is very suitable for solving classification problems, this opens the discussion for where else these techniques can be implemented to prove their diverse implementation, but more importantly if this dataset can be scaled up to include all Pineland Forests in NJ and make predictions about tree mortality for future forest fires. It will require a big effort to collect dataset across all forest lands but other data science techniques such as interpolation and kriging can be used to assist with field data collection.

V. Conclussion

We conclude that we were able to achieve high accuracy results that provide a window for future and possibilities of using machine learning techniques in forestry data, with bigger datasets better and more credible results can be achieved. By implementing necessary data preprocessing and oversampling techniques we were able to achieve a good working model with high accuracy results.

We received positive feedback from forestry service on this dataset which indicates a positive sign for future research applications.

Acknowledgment

Thanks to NJ State Forestry Supervisor William Zipse for providing the dataset and providing feedback on this project.

References

Brownlee, Jason. "Random Oversampling And Under sampling For Imbalanced Classification". *Machine Learning Mastery*, 2020, https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ [1]

"Categorical Encoding Using Label-Encoding And One-Hot-Encoder". *Medium*, 2020, https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd [2]