FINAL CAPSTONE REPORT

June 26 2020
By Sophie Ngo
Student# 88542

Your project problem statement - the underlying question you are seeing to answer or problem you are addressing i.e what are the goals of your project ?

The goal of the project is to detect edible and poisonous mushrooms. I was not able to detect multiple mushrooms that are either edible or poisonous. I, however, was able to differentiate one type of poisonous mushroom to one type of edible mushroom. The two types of mushrooms that I chose were the deadliest mushroom called the Death Cap and the most popular one called The Chicken in the Wood. I was still happy with the result because my ultimate goal is to identify the most deadly mushrooms. If someone ate a mushroom that caused him or her to be sick for two days, that would just equate to an unpleasant experience rather than their last experience.

I just simply did not have enough images for the model to learn, I also did not have the hard drive space to store them nor the computing power. I definitely could have used AWS to make those tasks a lot easier, but my budget could not accommodate it for this month. As for the images, I find there are not many images of one type of mushroom online. I want images of those mushrooms not only from every angle, but also when they are cut open in a certain position. I personally prefer to go out and take the pictures myself but it is currently not the season nor the right timing.

Background on the subject matter area of your dataset - why is this a good problem/subject area to apply data science techniques? How has it been addressed in the past?

I am passionate about mushrooms and I also know a lot of mushrooms. I am a part of the Mycology Society of San Francisco. It is a good problem to solve because if this model is deployed into an application, people could easily identify the different types of mushrooms and whether it is poisonous or edible. In conclusion, people will easily have more access to mushrooms. As a result, that will provide extra food and income to an individual, more mushrooms will help save the planet and meet the future demands for those who are vegans and vegetarians. On Kaggle, there are projects that identify which family does the mushroom fall under; it is a categorical image classification problem. There is also another interesting project on Kaggle, it detects poisonous mushrooms based on descriptive mushroom features.

The results of those Kaggle competitions are a lot better than I thought it would be. The image classification problem, one person managed to achieve 99% test accuracy and they used techniques that I have not yet encountered. As for the descriptive mushroom results, one person had 15 plus models trying to obtain the highest accuracy. The issue in this case

is that people's perception of certain descriptions are different. That might not be the best idea considering it could be life threatening.

Details on the source of the data and the dataset itself (including data format, structure and schema, etc)

The data is from Flickr, an image hosting service. I searched on Flickr using the Latin names of the mushrooms so there are minimum errors on the identity of the mushroom. The data was downloaded from Flickr. There are two types of mushrooms, one type of mushroom was poisonous, labelled poisonous and the other type was not, labelled non-poisonous. Then I created three folders, train, validation, and test and each one of those folders has a poisonous and non-poisonous folder with poisonous and non-poisonous mushroom images in them. The training folder has 70% of all the poisonous and non-poisonous mushroom images, the validation and testing folders have 15% each for poisonous and non-poisonous mushrooms.

A summary of the pre-processing, feature engineering and any other data cleaning/transformation, and exploratory data analysis(EDA) performed and the motivation and reasoning behind it.

For the pre-processing, I went through all the images (all ~ 4000 photos) and looked for any outlier images such as flowers or not that particular mushroom. Then those images are deleted. In the beginning, I was planning to have a few types of mushrooms in the poisonous and non-poisonous folders. I find the model was able to produce higher accuracy with less mushrooms. As a result, I only have two types of mushrooms in my model. I use my script called 'resize.py' to resize the image to an ideal height and width. The reason is because I want my pictures and its ratio to be relatively the same size so the pixels in each image are not drastically too big or too small compared to other images. The model would have an easier time to learn and the running time would be faster if the size of the images are similar.

Afterwards, I visualized the images in batches at random to see what images are going through the model under the data visualize category. Furthermore, it is a chance for me to see if I filtered out all the non-mushroom pictures. The images are labelled 1 or 0 in the y-train data set (1 for poisonous and 0 for non-poisonous). It is a bit different for the augmented model; the images of the mushrooms are visualized with rotation. Finally, I want to prepare my data with the proper input shape, thus it'll be ready for training.

A summary of all the modelling completed including the process of model evaluation, selection, and results.

For my first CNN model that I put together, I try with 1 to 3 convolution layers and I find that 2 convolution layers have the highest validation accuracy. Trial and errors identify the best numbers of filters and dense layers. When the model has 4 types of mushrooms each for poisonous and non-poisonous, the accuracy is lower roughly about 65-70% for validation and test opposed to two types of mushrooms which is approximately 80-85%. I added dropout to encounter any over fitting. In the future, I am going to train the model with one

mushroom at a time. As for the model evaluation, the confusion matrix is constructed using the test data set. I also evaluated the model with pictures showing the results of the predicted class and the actual results of the class. I know it is critical that my model properly predicts poisonous mushrooms because it could be life threatening. Therefore, I have to change the threshold of my model to it being poisonous if there is a 20% chance or more of it being poisonous instead of more than 50%.

I want to see if the model could learn if the images are augmented. Hence, this could help indicate whether the model overfits. If the model could not learn from the augmented images, then the chances are the original CNN is overfitting. The results show that the CNN model with augmentation has a validation accuracy of roughly 85-90% which means the original CNN model is probably not overfitting.

Finally, I wanted to compare my original model with the pre trained VGG 16 model. The results are the best compared to the other models. The validation and test accuracy is approximately 90-95%. For the VGG 16 model, I changed the model to sequential and changed the last dense layer to 2 from 1000 outputs.

Findings and conclusions based on all analysis and modelling of the data - how do your results compare against your initial goals & hypotheses?

My initial goal is to compare my CNN model to other models. I prefer to have a high validation and test accuracy due to the nature of my topic. My hypothesis was that the model is going to have a difficult time differentiating between the mushrooms because they all look like they have similar features.

My results show that the validation accuracy is higher depending on the model. I predicted that the CNN fine-tuned VGG16 model would perform the best and it turned out to be true. The reason is because it has 16 layers and it is pre-trained meaning that it has already proven to be successful with other data test set. It is a part of the Keras library too. My hypothesis came true, the features are just so similar that is why a quality model such as the VGG 16 is necessary for mushroom identification. My models end up with a pretty high accuracy score all across. Therefore, I am very content.

A final summary of the business applications of the project as well as potential next steps and future directions.

The model would have more potential if I add more images of mushrooms to my model. I want to add several thousand photos for each type of mushroom from all different angles and when they are cut open too. I want the end product to be an application with a front end, to get the pictures and a backend where the pictures are being processed in the model, then be output again in the frontend to depict the results. I will code the front end in Preact (smaller React), and the backend in Python Flask.