3 September 2020

Log book

Yesterday I talked to Dr. Garcia about how to format my references. I decided to use LaTeX as I am familiar with it. I have talked with Dr. Christian Jacob about my project and have sent his assistant (who is actually not his assistant) an email about when I could meet next with him. I prefer meeting with him next after I finish my research proposal.

I have created a document template in LaTeX but I haven't filled it with information yet. I have started using the biblatex package and have set it up. I plan to research my disease very soon, but I must first commit to a disease. Right now, I have 3 options. Kaggle is a good website.

1. Respiratory Sound Database
2. OSIC Pulmonary Fibrosis Progression
3. SIIM-ISIC Melanoma Classification

I am leaning towards option 1 because option 2 is about predicting lung function decline, not predicting who has the disease. Option 3 already has a lot of research put into it and similar methods are already in-use in hospitals.

September 5, 2020

Although I had started writing my paper about lung disease detection, my progress for today was invalidated. ☺ I had already found a good source about asthma diagnosis, had written an introduction, and a bit of the outline, I was told by my brother that he believes doing a project on Pulmonary Fibrosis would be better since it was something not a lot of people would have heard of.

For this reason, I have stopped my pursuit of my old project, and I am now editing my proposal for a project on Pulmonary Fibrosis. I am still writing the introduction. The whole proposal is due October 5, 2020, and must contain these sections:

✓ ① Working Title
✓ ② [Abstract]
✓ ③ Introduction
✓ ④ Objectives
✓ ⑤ Variables
✓ ⑥ Questions or Hypothesis
½ ⑦ Methodology
½ ⑧ Significance
½ ⑨ References

I will use this list as a checklist to gauge my progress.

September 9, 2020

The first thing I did this class was to give Dr. Garcia a copy of my proposal so far. I have not come close to finishing yet, but I believe it is a good idea to get feedback from her. For my bibliography management I am using biblatex and overleaf's in-built nature style bibliography style.

The majority of the information I have found is all from ~~google scholar~~. IPF* does not have any cure, is more common than ~~most sources~~ priorly anticipated, and * is very tough to determine the lethality of. Essentially the disease is when fibrosis occurs in the lungs for no known cause. Although there is a kaggle competition to determine lethality, I have to do research to see if this has been done before. ✓

Maybe try PUBMED rather than google scholar. See if you get good "review" papers —

*Idiopathic Pulmonary Fibrosis

September 11, 2020

I got Dr. Garcia's email reply yesterday, and I talked to her today. There were a couple things she outlined that I thought were important and that I should change.

✓ ① Primary sources. I should cite the primary sources and secondary sources rather than only citing secondary sources.
   @ I need to get a way to access the full version of papers. I could get my brother to get access to these

✓ ② I have been using google scholar for everything, but I could also use pubmed ☺
   @ I cited the Kaggle website when I shouldn't cite it as a source of info ✓

③ My citation style for Nature is inaccurate
   @ I should include (doi) as well

④ Expand my background research. I have to talk more about the disease
   ⓐ statistics
   ⓑ relevance
   ⓒ difficulties in diagnosis/treatment
   ⓓ how is it dealt with

✓ Also include info about ML used for other bio projects and how successfull it was and what models were used.

September 15, 2021

references

Over the weekend, I decided to change my paper style to science instead of nature. The science journal website has a section that contains a LaTeX template for anyone wishing to submit a paper to the journal. I was able to set up the environment after realizing that there is an error in the template.

By changing the style, and using the official science template, I should face no issues with citations. I also added the primary sources for a couple facts.

September 17, 2020

References

I found a good source for the Machine learning aspect of my project

① Reading Digits in Natural Images with Unsupervised Feature Learning

② High-Dimensional Pattern Regression Using Machine Learning from Medical Images to Continuous Clinical Variables.

One thing I noticed is that when I cite articles with google scholar, it doesn't give the doi, so I must go back and find this. That is what I will do today.

I found that there are many doi's available for medical papers, but there was one article I couldn't get the doi for:

"Automated quantification of ..."
F Maldonado, T Moua, et al. Year?

For now, I am searching for more articles about machine learning that I can use. The majority are about other diseases, or about image classification using features.

Sept 21, 2020

One thing I noticed is that the Science template doesn't include an option for doi in the citation. This means I have to create my own LaTeX bst file for bibtex. Since I struggled with that, I list the steps:

① Go to CTAN (Comprehensive Tex Archive Network) and
② Download the custom bib file/folder ✓
③ Open the folder and run the following

```
latex makebst.ins
latex makebst.tex
```
✓

④ Follow the instructions and answer the MC questionaire There is no need to install the full live-tex package. The question process is long and one I will go through later. For now I am using a simple bst file I made in a minute. ✓

Sept 23, 2020

Since the background research was done, today was quite easy. I worked through the objectives, variables, question, and didn't get time to work on method, but I looked at the Kaggle dataset for method inspiration.

Sept 25, 2020

Today, I explored Kaggle notebooks to see how each of them trained their models and what models they used. After this, I wrote the Methodology and Significance sections. The following methods were what were present in Kaggle

1. Simple Neural Network
2. Linear Decay
3. Linear Feature Engineering
4. Bayesian learning
5. Auto-Encoder training
6. QR with a CNN

There is one more model (not listed in Kaggle) that could be successful. This is using XGBoosting (extreme gradient boosting). I plan to create an ensemble of all these methods.

10/10

=> Excellent Notes for this months.—

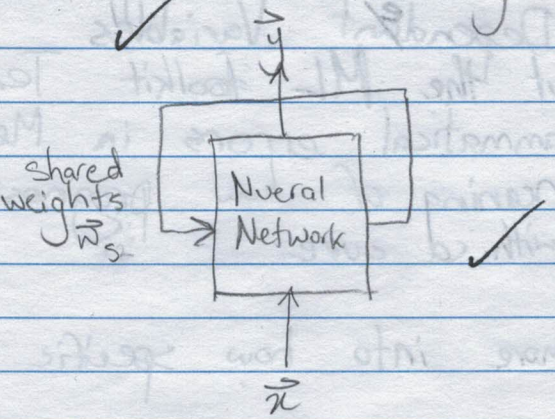
=> Do you have any notes about your meetings with Dr. Jacab?

Sept 29, 2020

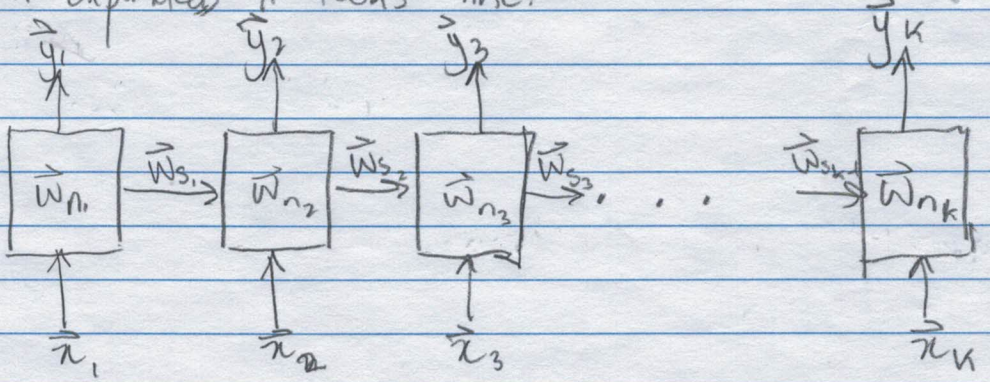
Yesterday I met with Dr. Jacob. We talked about my proposal and he gave me suggestions and edits on the pdf. I haven't taken a look at them.

He also told me to continue working on seeing the implementations of how other people have worked on the project on kaggle.

Today I watched some lectures about machine learning, namely the video about CNNs offered by the university MIT. The following structure is used:



When expanded it looks like:



Oct 6, 2020

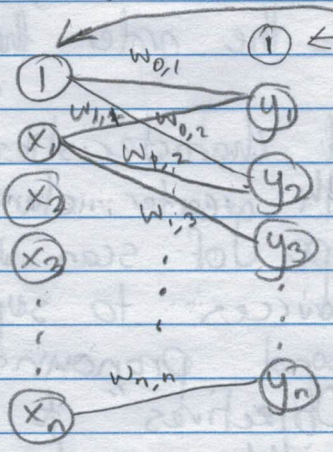
Today I made edits on my proposal. The following were the edits:

- Changing Dr. Jacobs title ✓
- Changing the objective of the project ✓
  - ↳ The dataset only allows for the prediction of the next 3 measurements given a change in time
  - ↳ Added web links to kaggle sources ✓
- Changed Dependent Variables
- Wrote about the ML toolkit Tensor Flow.
- Changed grammatical errors in Methodology ✓
- Clarified meaning of how prognosis helps doctors come up with a cure.

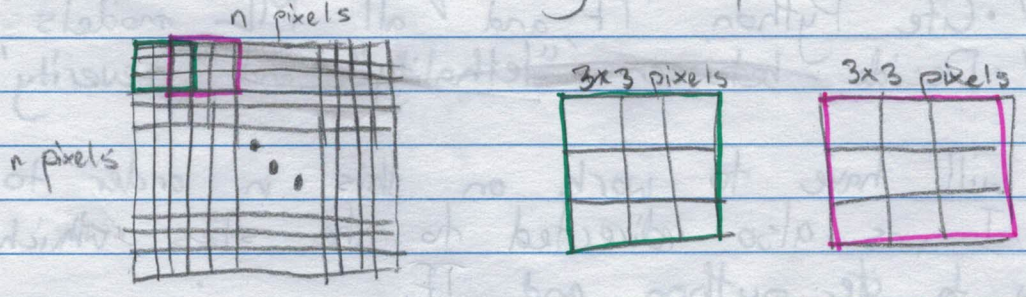
I will look more into how specific methods work. ✓

9 October 2020

Today I took a look at the ML structure. Each layer is made of perceptrons (neurons) and they connect with a weight to the next layer.



There is a 1 so that each layer also has a bias.
The CNN uses the following as inputs



Each 3x3 is an input to the network.

Pros	Cons
- The network knows its surrounding pixels ✓ - Works with images ✓	- Very computationally expensive ✓

14 October, 2020

Dr. Garcia sent back my draft of the proposal with many suggestions. Here are some:

- Citation are numbered in the order they are called in the text
- The effects and characteristics of IPF should be explained with greater clarity
- Specify the type of scan used
- Provide more sources to support claims
- Don't use personal pronouns
- Move some "Objectives" to "Methodology".
- Spelling of variables such as independent, etc.
- Move location of Question
  - Include IPF in question statement
- Cite Python, TF, and all ML models
- Decide between "lethality" and "severity"

I will have to work on this in order to fix it all. I was also redirected to the sites which describe how to cite python and TF.

October 16 2020

Spent fixing suggestions
Created .bst file to style citations

October 20 2020

Submitted paper after all suggestions fixed

October 22, 2020

great!

Today I met with Dr. Jacob again, but since the deadline for the proposal hasn't yet come, there wasn't too much to talk about, and so the meeting was short.

I also got back advice from Dr. Garcia about how I could improve my presentation.

- Talk about how machine learning has proven to work for something like this in the past
  - ↳ Maybe a different disease
  - ↳ How was it different than what you are doing currently?
- Is a CT scan a good descriptor of the severity of IPF?
- Provide citations to "k-fold" validation learning
- Explain why I can't choose the loss function right now in the dependent variables section
- ⇒ No more "I" or "we"
- Cite all ML methods (I forgot to before)
- Fix a References styling (to do with paper number)

October 26, 2020

Today is the day of the oral presentation. I have fixed up the slideshow to include all elements of my proposal. I also got a suggestion by Dr. Garcia that my significance part is good but there could be one improvement.

- Include and repeat stats about how IPF affects the population and how prevalent it is.

Other than this, I just spent the time reciting the presentation.

Oct 28, 2020

Today follows the presentation. There are some things left for me to do before I submit my final proposal. This includes:

- Explain why IPF is important
  - ⇒ How many people does it affect

The presentation went well, and I think that I should now begin to make the first model. Here is the order of models, I will make

1) Linear Decay
2) Linear Feature Engineering ✓
3) ~~X~~ Boosting
4) Bayesian Learning
5) Auto-Encoder Training
6) QR with CNN

in that order.

I also talked to Dr Garcia today and she said that I can improve my presentation ✓

- Don't include figures you won't talk about ✓
- Include more text on the slides
- Q: Why is my method > CALIPER ✓
- Increase presentation organization ✓

Great job keeping the logbook this month!

10/10

10 November 2020

I finished up the linear regression model that I have been working since 8 November 2020. Here is the outline of the program:

Import neccesary libraries
↓
Read training data
↓
Modify data frame (add extra columns)
↓
Split data into training and testing
↓
Make features to be used by tf (including derived)
↓
Make the input functions for the model for testing and training
↓
Make the model
↓
Train the model
↓
Test the model

The problem with linear model, though, is that:

1) There is no way to get a measure of confidence (required for metric of effectivity)
2) The average_loss is 46702.13 (huge)

As we can see, the linear model is inefficient, so I will make a neural network now.

November 11, 2020

I had emailed Dr. Jacob about having meetings every 2 weeks and I sent him the final version of my proposal.
We will now meet every 2 weeks on a thursday. This will start from the 19th of November, 2020, and will be until December 17th for sure.
Here is the list of all tasks I need to do.

Task	Estimated time	Rough deadline	Date Done
EDA	3 days	Nov 18, 2020	Nov 14, 2020
LR	3 days	Nov 11, 2020	Nov 11, 2020
SNN	5 days	Nov 23, 2020	Nov 24, 2020
Lin Dec	4 days	Nov 27, 2020	
LR w confid.	4 days	Nov 15, 2020	Nov 20, 2020
EGBoost	4 days	Dec 24, 2020	
Bayesian	6 days	Dec 3, 2020	
Auto-encoder	5 days	Dec 15, 2020	
QR w CNN	7 days	Dec 10, 2020	
tuning mloss	3 days	Dec 27, 2020	
Ensemble	5 days	Dec 20, 2020	
Transfer learning	5 days	Jan 1, 2020	

November 13, 2020

I started work on the EDA instead of LR w confidence. I am using the kaggle notebook "basic EDA" as reference. Here are the graphs I will make ✓

- ✓ - Distribution of checkup weeks
- ✓ - Distribution of FVC quantities
- ✓ - Distribution of patient age
- ✓ - Distribution of patient smoking status
- ✓ - Distribution of patient sex
- ✓ - Distribution of percent values
- ✓ - Scatterplot of FVC vs percent
- ✓ - Scatterplot of FVC and percent vs weeks
- ✓ - Violin plot of FVC and percent WRT sex and smoking status
- ✓ - Scatterplot of FVC and percent vs age

I do realise that I should also analyse the DICOM format CT scans, but this is for later. ✓

November 15, 2020

Today I finished off the list of graphs, but I didn't start the DICOM analysis yet. I save this for later. ✓

November 16, 2020

I found an example of a linear feature engineering model that was able to produce a measure of confidence, so I am now following their steps to do the same.

train, test, and submission dataframes are combined into one for manipulation.

Normally we have:

train has features:

patient, weeks, FVC, Percent, Age, Sex, Smoking Status

test has features:

patient, weeks, FVC, Percent, Age, Sex, SmokingStatus ✓

submission has:

patient_week, FVC, Confidence

Now the combined dataframe can be manipulated. The following features will be used in the models

Feature	Source	Normalized
Age	Age	y
First Week	Week	y
First FVC	FVC	y
First Percent	Percent	y
Weeks Passed	Derived	y
Height	Derived	y
Female	Sex ↓	y
Male		y
Currently Smokes	Smoking Status ↓	y
Ex-smoker		y
Never Smoked	↓	y

Note that the Smoking Status and Sex features are split into their constituent values, and then all features are normalized.

November 19, 2020

Today I made the Huber Regressor model and fitted it to the data linearly. After this, predictions were made and I made plotly graphs of:
- coefficients of the model (how each feature changes FVC)
- predictions vs. actual FVC
- Error

And I calculated:
- mse and mae metrics
- laplace log metric

November 20, 2020

I submitted the linear model to Kaggle and noticed one thing:
- Internet is not allowed so I had to remove the graphs in order to get the competition metric (I got a score of -6.9)

What does this mean?

November 21, 2020

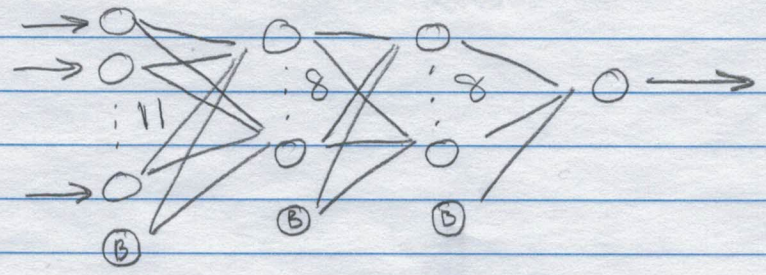
I reused the data from the LR and I just built the SNN above it. I am following a tutorial and noticed one thing. As I graph loss, mse, and mae, I notice a downward trend as the model trains, but when I graph accuracy with the epochs (how many times the model trains), the accuracy is consistently 0.00%

✓

November 23, 2020

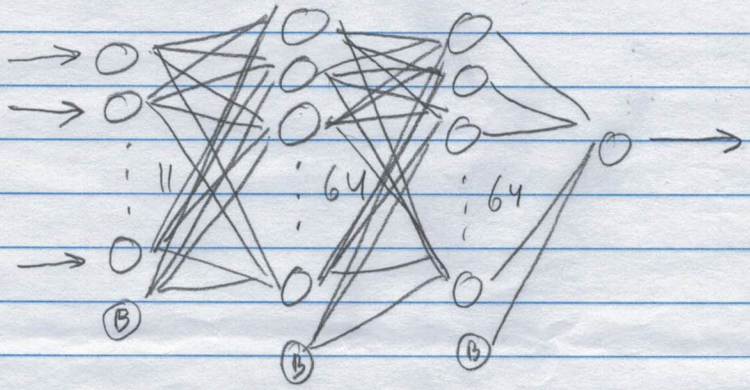
Today I figured out why accuracy never increased. It is because accuracy is only a metric for categorical model, not predictive models. I also then tried experimenting with model attributes.

Model Architecture



11×8×8×1

✓



11×64×64×1

✓

I also experimented with the activation function too:

Sigmoid and Relu functions

I used dropout methods to avoid overfitting with a droupout percent of 50%.

10/10 Excellent progress this month, Anna!

Remember Lit. review (Introduction) is due Dec 1st.

November 28, 2020

The introduction to my paper is due December 1st so I decided to start that today. I recycled a lot of the introduction to my proposal but made some wording changes. I also lost citation points in my proposal because of the presence of a journal number, so I will make a new .bst soon.

November 30, 2020

Today I am making the .bst file. Here are the answers.

```
* * ScienceDOI, Science, * * * * * * * * * * c, *
i, * *, l, y, y, * * *, p, x, * * * * * * * x, * * b, c, b,
s, * * * * * * * * * * * * * * * * * * * x, * * * *
*, *, *, *, a, a, a, * * * * * * x, *, i, * * u, *, i, *, y
```

Still, I am getting the same problem as before in that there is excessive information in the citation. For this reason, I take the ScienceDOI.bst and modify the Science.bst. This takes longer than anticipated. Here is the gist.

Move the following code:

1. FUNCTION {format.doi}
2. Add doi entry
3. Change the Article class to include doi

December 1, 2020

I submitted the introduction section today after fixing up the wording of the question statement.

December 3, 2020

I was returned the introduction section with edits. I removed the abstract section, made some edits (with spelling), and re-submitted the paper.

December 8, 2020

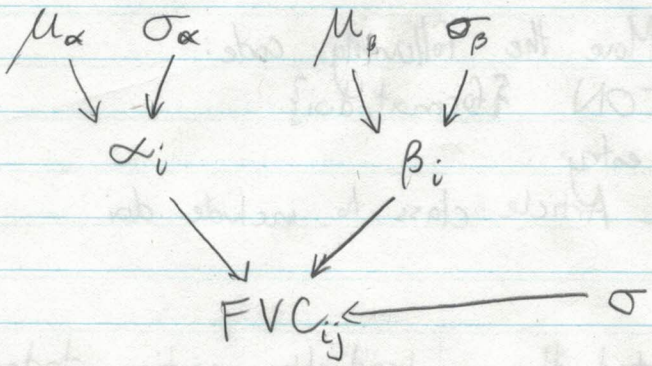
The method of bayesian learning is very different from regular NNs. Here are some interesting things with how it was implemented by Carlos Souza

- No baseline FVC and Percent measurements
- Keep the patient field
- Approach this as a matrix completion task

This relies on a principle called Ockham's razor which is supports the idea that simpler is better. Here are the features:

- FVC
- Week
- Patient ID

The method assumes a linear relationship and has a intercept (α) and a slope (β). These α, β can be different for every patient. They are still related though. So not all α β are the same (pooled model) but they are not independant either (unpooled). It is partially pooled. See the digram where α_i, β_i are coming from a common distribution.



i is the patient, j is the week. $FVC_{ij} \sim N(\alpha_i + t\beta_i, \sigma)$

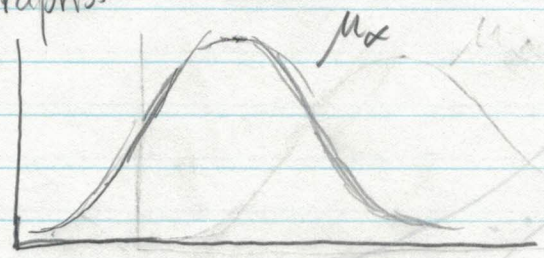
$\mu_\alpha \sim N(1700, 400)$ $\quad$ $\mu_\beta \sim N(-4, 1)$

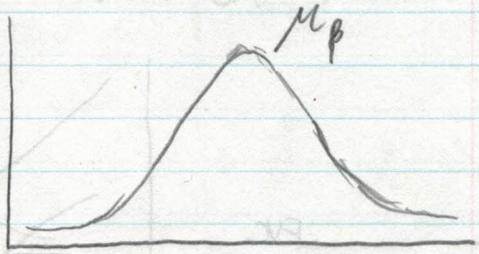
$\sigma_\alpha \sim |N(0, 1000)|$ $\quad$ $\sigma_\beta \sim |N(0, 5)|$

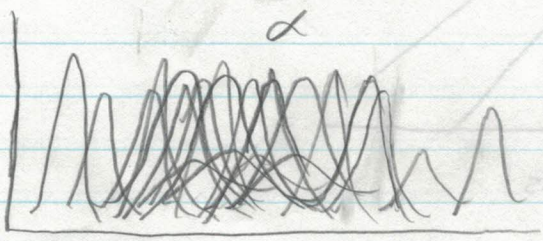
$\sigma \sim |N(0, 150)|$

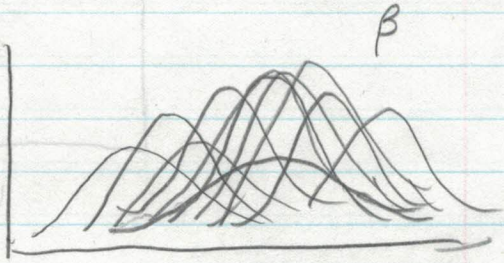
December 10, 2020

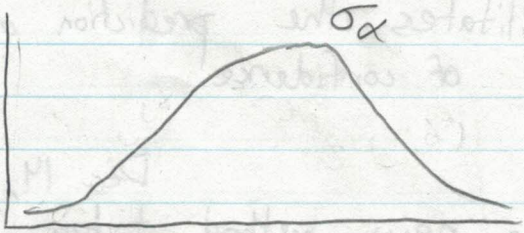
After training the Bayesian model, one obtains many distributions. We know μ_α, μ_β, σ_α, σ_β, and α, β are approximately normally distributed. Here are the approx. graphs.

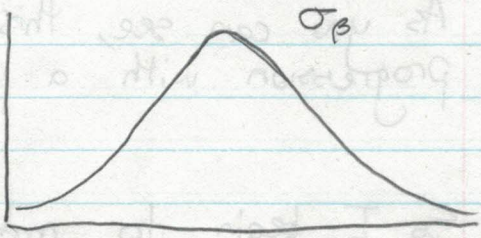


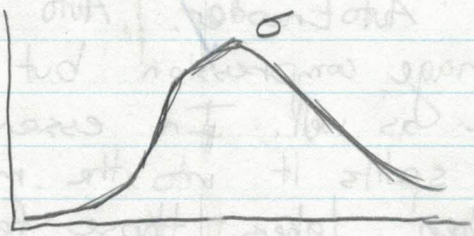








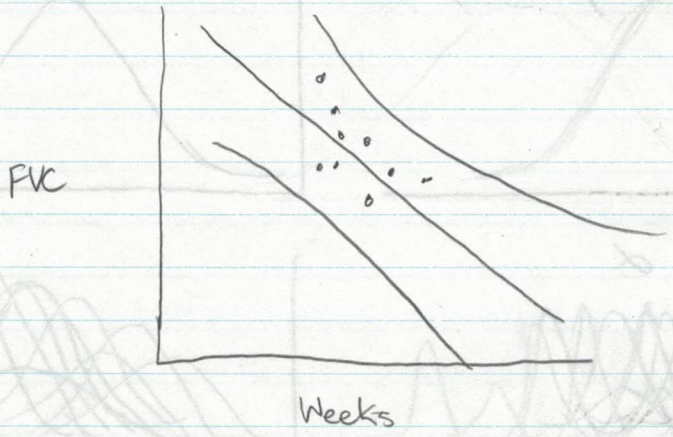




We can see many distributions of α which show that α and β are partially pooled.

Dec 12, 2020

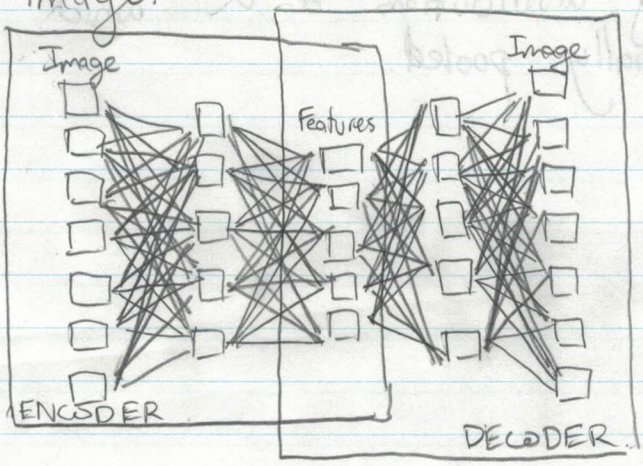
It turns out that after training the Bayesian model, an interesting graph is created that corresponds to the ~~linear~~ confidence. Here, the wider the section, the less confident the model.



As you can see, this facilitates the prediction of the progression with a level of confidence.

Dec 14, 2020

So I begin to make a new method today. This method is called an AutoEncoder. Auto Encoders are generally used for image compression but can be useful for this purpose as well. In essence, the encoder takes an image and splits it into the most important features. A decoder then takes those features, and reconstructs the image.



Even though I only care about the encoder, I will still have to use the decoder for the traing process. Luckily I found Kaggle User Welf Crozzo to have a TF encoder already built, so I can just use the premade encoder to generate more features to use in a SNN

Dec 16, 2020

I copied over the code from Welf Crozzo's notebook over to my notebook and copied the SNN code after it. The one think I have to do is to add all features to the cumulative dataset. The code I made to do this roughly does the following:

1. Get all unique patients
2. Look up all of their DICOMs
3. Convert the DICOMs to features
4. Add those features to all instances of that patient

But I get an error when running the code, and retraining the AutoEncoder everytime takes arond 20 to 30 mins.

Dec 19, 2020

I realize that Kaggle allows me to store files so I only have to train the encoder once, and load it again if I ever need it. This is quite usefull for troubleshooting my error, so I use this. I am not exactly sure where the error stems from, but I know that it happens when I try to add the encoder generated features.

Dec 20, 2020

Also as a note, I explain the Laplace Log Likelihood here as I forgot to explain before. I use the Laplace Log Metric to measure the accuracy of the models. I recently read another notebook explaining this. Laplace Log Likelihood is a function of the actual values, the predicted values, and the confidence. Interestingly, a lower confidence value means more confidence.

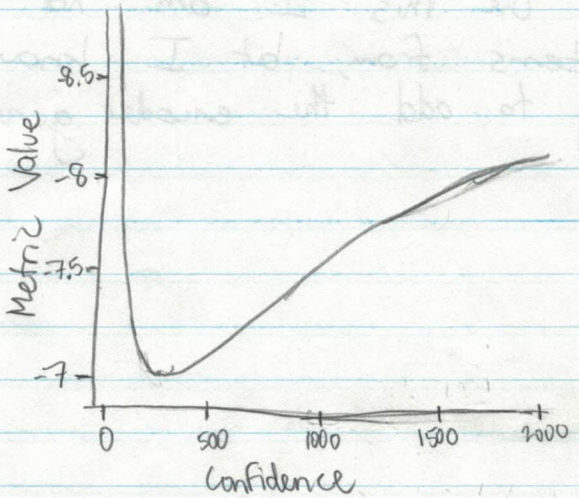
When the most basic model is implemented (predicts the average of FVC), then a LLL of -8.02318 is attained, so any model which performs worse than -8.02318 is useless.

If the confidence is σ, then here is how the metric is computed:

$$\sigma_{clipped} = \max(\sigma, 70)$$
$$\Delta = \min(|FVC_{true} - FVC_{predicted}|, 1000)$$
$$\text{metric} = \frac{-\sqrt{2}\Delta}{\sigma_{clipped}} - \ln(\sqrt{2}\,\sigma_{clipped})$$

This means that if the model shows extremely poor confidence or is extremely inaccurate, it is not overly reprecussed. This also means that the worst possible score is -24.798 After the metric is computed for each datum, the avg of all metrics is then taken.

The following graph shows (for a predicted FVC of 2700 and a real FVC of 2500) that a confidence too high or low worstens the LLL.



Dec 22, 2020

Today I continue working on the AutoEncoder. I have noticed that the model I saved last time has dissapeared. This means I had to train my model again. This makes me inclined to create a local copy. I have created a github for the project as a result (a private repo) and have added my papers. Here is the file heirarcy

OSIC-IPF
- papers
  - cited_papers
  - paper
  - proposal
- models
  - linear_regression
  - simple_neural_net
  - autoencoder

Dec 24, 2020

After copying over the papers into the github and creating a .gitignore, I created a conda environment with all the python packages: (version python 3.7)

- pandas
- opencv
- matplotlib
- numpy
- tensorflow
- tqdm
- pydicom
- scikit-learn

I then proceeded to copy over my linear regression model to the github. I made sure to split the program into different sections where each would write to a file. Here is the heirarchy for each model:

linear_regression
- extras
  - model.sav
  - formatted_data.csv
- create_model.py
- format_data.py
- make_predictions.py
- submission.csv

Dec 25, 2020

Today, I transferred over the simple neural network model in a similar fashion. There were some more difficulties I encountered along the way, but I resolved them.

Dec 26, 2020

Today, when making the autoencoder, I noticed that it would not be feasable to run it locally so I ran it on kaggle and simply imported the model.h5 folder. I will have to run the model on kaggle as well.

Dec 30, 2020

Today I was finally able to debug all the problems I was experiencing and got format_data to work. Here are the features of the data:

- Age } scaled, direct
- FirstWeek
- FirstFVC
- FirstPercent
- WeeksPassed } scaled, derived
- Height
- Male
- Female
- Currently Smokes
- Ex-smoker } binary, derived
- Never smoked
- f0
- f1
- :
- :
- f1919 } scaled, derived from encoder

Excellent Notes, Amar

10/10

Jan 2, 2021

great!

I completed the neural network that takes the features and makes predictions. The one problem I see is that when compared to the training data, a competition metric of only -7. is produced. This is worse than the SNN. II will experiment with different structures for the neural network.

Jan 3, 2021

Today I revisited the SNN, changed around the structures and got an even greater accuracy. I found that the most effective SNN architecture might not be the 11×64×64×1 architecture or the 11×8×8×1, but after trying a 11×1, 11×10×1, and many other architectures, I have settled with the 11×12×1 architecture which allows me to keep a low dropout rate (since there is less fear of overfitting). Comp metric: -6.6715

Jan 6, 2021

I tried a similar method as the SNN to try and reduce the comp. metric for the autoencoder. I tried the 1930×1830×100×1 structure at first, but after remembering that the SNN performed better with a simpler architecture, I tried a 1930×1830×1 architecture. I found that the results were still worse than the SNN, so I assume it simply requires more training epochs.

Jan 9, 2021

Today I learnt of lazypredict. This library tests multiple ML models on the data all at once. This is revolutionary because many of the models I wanted to implement such as gradient boosting and XGBoosting are built in. The following are the regression models:

- Gradient Boosting
- XGBoosting
- Poisson
- Extatree

- Random Forest Regressor
- LGBM Regressor
- Extratrees Regressor
- Bagging Regressor
- Hist GB Regressor
- Decision Tree
- KNeighbor
- RANSAC
- TransformedTarget
- Linear
- Adaboost

The question now remains what the remaining models are.

Jan 13, 2021

I started implementing the lazy regressor last class and continue today. The conda environment installer couldn't install lazypredict, so I had to use pip.

Since lazypredict is not something that predicts stuff based on it's own code, more packages must be installed such as xgboost, lightgbm, pytest. I ran the program and got results.

Here are the rankings (Only top positions)

Model	R-squared	RMSE
RandomForest	0.97	0.02
LGBM	0.97	0.02
ExtraTrees	0.96	0.02
HistGB	0.96	0.02
XGB	0.96	0.02
Bagging	0.96	0.02
GB	0.96	0.03
DecisionTree	0.95	0.03
NuSVR	0.94	0.03
ExtraTreeRegressor	0.94	0.03

All of these performed better than linear regression so I think these might be the best.

Jan 15, 2021

Going back to the autoencoder, I move everything back to Kaggle in order to get the testing set Laplace Log Metric reading. Although I thought it would run quite well, saving the notebook and then submitting it takes forever. Saving the notebook alone takes around 40 to 50 minutes, and running it (after submitting) gives a timeout notebook error.

Jan 18, 2021

I started by realizing something. I have already done the tough conversion from the AutoEncoder to tabular data, so I could go ahead and just implement a LR ontop of it. That is what I did today and got surprisingly good results. When I did the LLL with the training data and got a score of -6.348. It appears that this is better than most scores. Similarily though, it gives the timeout notebook error when submitting to the Kaggle competition.

Jan 20, 2021

I have a meeting with Dr Jacob coming up, and I prepare to show him what I have done. I copied over the starter notebook for the quartile regression and convolutional neural network. I don't understand how the convolutional network works so I have to research that.

Jan 22, 2021

From what I understand, the OSIC multiple quantile regression starter notebook uses a convolutional nueral net on the tabular data (not the images) and uses quantile regression to determine the uncertainty based on the deviance from the ground truth.

Michael Kazachok also created a variation of this notebook that uses the images using Decay theorey. The model uses a CNN to predict the coefficients of the following decay theory:

$$FVC = a.quantile(0.75) \times (week - week_{test}) + FVC_{test}$$
$$Confidence = Percent + a.quantile(0.75) \times abs(week - week_{test})$$

Where a is what is predicted.
Contrary to belief though, the score recieved with both methods is very comprable.

Jan 26, 2021

Today when I met with Dr. Jacob, I was given two main suggestions
1) Make my own testing set to test the AutoEncoder based models.
2) Ensemble all the methods.
Other than this, he stated that everything looks good. The first recommendation was quite unfeasable, but I think I will implement a simple ensemble before the end of the month.
I originally delayed the ensemble method because I was unsure how I was going to implement it, but I think I have an idea. ☺

Jan 28 2021

I will make the ensemble as a statistical method that first runs all the individual methods to get their submission.csv files then averages them somehow. I made the part that gets all the submission files today.

Fantastic Progress!

10/10

Jan 31, 2021

Although I began work on the Ensemble method, it doesn't seem to be feasable. For one, it is very hard to get the models to all make predictions of the same format. Here are the issues.

1) It would give notebook timeout error
   ↳ this means that the private testing and public testing metrics are unavaliable
2) It is tough to get the model predictions of the ~~testing~~ training data and then combine it

For this reason, since it would take ages to fix any bug, I'll leave out the ensemble method.

Feb 3, 2021

Since I am not doing the ensemble method, I can begin to enter stuff into the CYSF portal. Dr. Garcia set a portal check-in on the 10th of Feb to complete the background and proceedure sections. The majority of my work on the background section seems easy as I just have to transfer the background of my paper into the portal as point form (to make it understandable and easy to read). I also keep the citations.

Feb 5, 2021

I have been experimenting with the CYSF portal and do not find it too easy to use. I was able to get images into the portal though, which means that I can start importing graphs. Today I finished the introduction.

Feb 8, 2021

I work on the methods today, listing out every method I used and making an enumerated list of steps to make the model.

Feb 10, 2021

I met with Dr. Garcia and got the following recomendations.

- Add figure numbers and captions
- Include a flowchart in the methods section.

So I made a flowchart with google draw which contains all the models, and I added it to the portal. I also added figure numbers and image citations.

Feb 12, 2021

The portal check-in for the results and conclusion section is the 17th, and I won't have much free time as I qualified for the repêchage which starts soon. Today I begin work on the analysis section. I first rerun all the models, take screenshots of the graphs, and input them into the analysis section under their respective model.

Feb 17, 2021

I showed Dr. Garcia what I have thus far, and realize I have to do a bit of work over again because the graphs produced with python do not have units on the axis and also do not have titles. I spent some time today putting up axis labels and titles by changing the python code. I also have a presentation on Feb 23rd that is meant to simulate the video.

Feb 22, 2021

Today I made the powerpoint for tomorrows presentation and fixed the labels on all the graphs.

Feb 23, 2021

After reciting the presentation (which was 5 minutes overtime) I got certain feedback.

- Remove unnecessary data exploration
- Don't explain all models
- Emphasize the importance
- [illegible]

So I will try to implement the feedback.

Feb 25 2021

The portal closes on March 4th for the school sciencefair which means that I have to get everything done by then. I have made edits on the original presentation I had, and sent it to Dr. Garcia for investigation.

Feb 27, 2021

I worked in the portal, adding information about LLL to the methods, and adding some EDA graphs.

Feb 28, 2021

I added results graphs to the analysis section today and did some refinement on the ppt to address the topic better.