**DAEN 690 Presentation April 2, 2024**

**Slide 66 – Machine Learning – ISSC Volume Predictions**

* A big focus of our project has been on the data visualization side for analysis of past ice supersaturated conditions which can be helpful for climatologists and meteorologists. However, secondarily, our client wanted a predictions dashboard, which meant creating some type of machine learning model that could potentially predict if an ice supersaturated region would form on a given day, which would allow airline companies to adjust their flight plans accordingly to avoid these ice supersaturated regions.
* To achieve this, we scoped our goal to predict the volume of ISSC for the next calendar day, using a time series model to forecast these future days. A day with higher volume levels would indicate a day that would likely experience ice supersaturated conditions and if the volume is high enough, could explain the vertical depth and number of hours the conditions would occur.
* Our end goal is after the modeling is completed, to integrate the prediction model into the predictive dashboard.
* In these next couple slides I’m going to go over the two machine learning models we built and evaluated, which were the Autoregressive Integrated Moving Average model and the Long Short-Term memory model

**Slide 67 - Machine Learning – Preprocessing/Feature Engineering**

* Before going into the two models, I want to first talk about the various preprocessing and feature engineering aspects that we did.
* For preprocessing we
  + Filtered the data to the years 2022 and 2023, which helped to reduce the model complexity as well as more recent information could be a better predictor for future events as our time prediction window is shorter than long term.
  + Converted the ISSC column character to integers to have a proper dependent variable to feed into our model
  + Made the date the index as this was needed for our time series models
  + Removed extreme outliers. We had one day that experienced 55 instances of ice supersaturated conditions, which in initial testing really threw off our models as the mean was around 2 instances for days with ISSC.
* For Feature Engineering we created and tried a variety of different features in our model
  + Originally, we created a volume feature which basically calculated the number of instances per day ice supersaturated conditions occurred and that would be our dependent variable.
  + Our independent variables we created after that were lag variables for volume of ISSC per day.
  + Then we tried the mean daily temperature and relative humidity to ice, which improved our model predictions more
  + From there we thought maybe try creating an improved version of those two features and created a temperature daily volume feature which represents the number of times per day the temperature is below -42 degrees Fahrenheit and the relative humidity to ice daily volume feature which represents the number of times per day the relative humidity to ice is above 100 percent.
  + We tried other features like rolling average and exponential smoothing but had little success with those features.

**Slide 68 – Machine Learning – ARIMA Model Evaluation**

* After doing some feature engineering, we evaluated the ARIMA model and LSTM model.
* For the ARIMA model, we first built the model and then created a parameter tuning function which basically tried various combinations of P, D, and Q values.
* P represents the number of lag observations in the model
* D represents the number of times the raw observations are differenced
* Q represents the size of the moving average window.
* When running our parameter tuning function, we were given 5, 0, 2 as our optimal parameters.
* When plugging these parameters into our ARIMA model, we were left with a mean square average of 0.988 and the following Actual versus predicted plot, which as you can see is underwhelming.

**Slide 69 – Machine Learning – LSTM Model Evaluation**

* Moving on to the LSTM model, in our initial testing we found better success and so more of the effort was taken on optimizing this model.
* To fully optimize this model, we first evaluated the optimal number of lag variables for ISSC Daily Volume, Temperature Daily Volume, and Relative Humidity to Ice Daily Volume.
* After creating a function to run the different combinations, we found that ISSC Daily Volume of 2, Temperature Daily Volume of 0, and Relative Humidity to Ice Daily Volume of 1 was the optimal number of lag features for each of our variables.
* After optimizing the lag variables, we went to optimizing the hyperparameters.
* We tested different values for units, learning rates, and batch sizes.
* We visualized the different combinations in a heatmap shown on the right-hand side.
* The optimal hyperparameters were units set to 100, learning rate set to 0.01 and batch size set to 64.

**Slide 70 – Machine Learning – LSTM Model Evaluation**

* As a result of the lag variable testing and hyperparameter tuning, we trained and tested a newly developed model that was more fully optimized. As shown at the bottom, the left-hand side was our model’s performance before the optimization and on the right is our optimized model.
* Although we saw a great improvement, limitations persist that will need be researched in future projects, such as
  + Addressing the missing values from an incomplete radiosonde dataset
  + New research into understanding the unpredictability of relative humidity to ice
  + And additional modeling time with additional computing resources to dive even deeper into the feature engineering and parameter tuning.

**Slide 71 – Risks and Planned Mitigations**

* As for risks and mitigations, very similar issues still persist with trying to keep all programmers on the same page as well as keeping communication channels open with the Tableau developers.
* Additionally, as we are nearing the end of the semester, trying to keep on top of the timeline of the project.

**Slide 72 – Weekly Partner Meeting**

* This week we will be meeting with Dr. Sherry this Thursday. We will be going over our Tableau dashboard as well as go over the modeling results and get his thoughts on both before heading into our last Sprint.

Units is the number of neurons connected the neural network layer