STA325 Case Study

Chase Mathis, Dillan Sant

Introduction

Nobel Prize winning Physicist Richard Feynman said that turbulence was "the most important unsolved problem of classical physics". Although, we still unexpectedly feel bumps on airplanes and have trouble truly predicting where a Hurricane will strike land, we hope to add to the field of research studying how to model turbulent systems. We position our research in two key areas:

1. Prediction
2. Statistical Inference

Prediction is important in the context of this problem as data from turbulent systems is very computationally expensive to obtain. Therefore, experiments using data on the evolution turbulent systems need better, quicker, cheaper methods to get data.

Statistical Inference is also important as it can help elucidate the connections between turbulence and initial settings. We hope to build a model that gives clarity on this as well.

Methodology

We are given 112 data points, but we only use 89 data points for training our model as to not overfit our model. We will examine numerous different types of models including:

- A Simple Linear Regression Model
- More Complex Non-Linear Regression Model
- Tree Based Model
- Boosted Tree Model

In addition to our splitting of data for training and testing, we further split our training data where 66 data points are used for training and 23 data points are used for selecting our best model. When fitting the models mentioned above, we benchmark their predictive performance on the validation set.

While cross-validation is a common technique to estimate the true test-error, we believe that it would be too computationally expensive to fit many different folds of all of the models we wish to predict. This is a shortcoming, and provides for future investigation.

The data has `4` response variables and `3` input variables. Each output variable is a raw moment of a final turbulent distribution. We've converted the raw moments to central moments so as for better interpretability. We keep the first raw moment, however, as the first central moment is 0.

As there are only three input variables, we will stray away from methods which penalize models with high numbers of correlated predictors i.e. (lasso/ridge regression). Our model is already sparse so there is there is no need to make it more sparse and there is low risk that the predictors are highly correlated.

Quick Transformations for our Input Variables

Reynolds Number	Count	Froude Number	Count
90	27	0.052	23
224	26	0.300	20
398	13	Inf	23

Reynolds and Froude numbers are non-continuous as there is only three buckets for the numbers. Therefore, we are going to make them a categorical variable to help interpretability.

Simple Linear Regression Benchmark

Table 2: Mean Squared Error using Simple Linear Regression

Mean	Variance	Skewness	Kurtosis
0.0002482	33003.25	2.128693e+12	1.537524e+20

It is clear looking at the MSEs of the moment that the using a Simple Linear Regression model to predict anything but the `mean` is not a good idea as the performance is not good, with increasingly high MSEs for each moment. Nonetheless, it serves as a good baseline for assessing further complex models.

Polynomials

We conducted cross-validation to determine the best fitting degree-polynomial for Stokes number. Our chosen polynomial models will include some degree polynomial of Stokes number, while adding Froude and Reynolds number as well. Since these variables are factors, we cannot make them polynomials, however, we still include these variables in our model since we decided against sparse models since we only start with three predictors, and all three predictors have a significant effect on the first four moments of particle cluster volume.

The optimal degree polynomials for predicting mean, variance, skew, and kurtosis are all 2 for Stokes Number.

Using 2nd degree polynomial models while also predicting the moments with Froude and Reynolds numbers, we get lower MSEs for all four moments, however these MSEs are still relatively high, indicating that using polynomials would also not be the best way of predicting the moments.

Interaction Effects

We check to see if there are significant interaction effects between the three predictors, and then construct a model with these interaction effects. We expect there to be significant interaction effects between the three predictors due to the similarity between the three numbers. For example, we expect Reynolds number (fluid turbulence) to have some relationship with Frouds number (gravitational acceleration).

Trees

Finally, we construct a tree-based model and a boosted model.

We tested whether pruning our tree would help predictive performance, but found that it did not. Therefore, we left all our trees as fully grown trees.

Results: Evaluating Our Model & Investigating Relationships

Evaluating Our Model

```
[1] "Tree Based Model" "Tree Based Model" "Tree Based Model" "Tree Based Model"
```

Our final results of the MSE for each model for all of the four moments leads us to selecting the tree-based model. This is the model that has the lowest MSE for `mean`, `variance`, `skew`, and `kurtosis`. The tree-based model makes for interpretable model for a tricky set of data which included only one quantitative predictor (Stokes number) and two factor variables (Reynolds number and Frouds number). While the data supports us using a tree-based model, we understand that there comes some uncertainty with this model selection. We randomly sampled a validation set to test our error and if we sampled a different batch of data, a different model may have predicted better.

Investigating Relationships

For each response variable, the input variables have a different effect if we examine the regression output. For the:

- Mean
  - As the size/density of the particles increase, the mean of the distribution increases as well.
  - A medium/ high gravitational acceleration is associated with a lower mean
  - A medium/high turbulence is associated with a lower mean.

  Mean is difficult to interpret as it is subjective based off of what coordinates one is using. Therefore, the other moments will be more exciting to find connections between.
- Variance
  - As the size/density of particles increase, the variance of the particle clustering distribution increases as well. This makes sense as larger/more dense particles are less affected by gravity and other forces and are more associated with their beginning positions.
  - Medium/high turbulence is associated with lower variance of the particle clustering distribution. This also intuitively makes sense as when there is more turbulence, particles are more likely to hit each other resulting in more similar interactions between particles. As the law of large numbers claims, large numbers of interactions should move particle distributions towards the mean thus reducing its variance.
  - Interestingly medium gravity decreases the variance of the particle clustering distribution, while infinite gravity also decreases the variance but not as much as compared to medium turbulence. This may be a product on what exactly "infinite" gravity is and how it is used in the numeric solvers to get particle clusters.
- Skewness
  - As Stokes number (the particles' size/density) increases, we expect the skew of the spatial distribution and clustering of particles to increase.

  - As Reynolds number (fluid turbulence) increases from low to medium, low to high, and medium to high, we expect the skew of the spatial distribution and clustering of particles to decrease. Scientifically, this makes logical sense because higher fluid turbulence would cause the particles to interact more and draw more closely to the mean, ultimately decreasing the skew of the distribution.
  - Moving from a low Froude number to a medium or infinite Froude number, we expect the skew of the spatial distribution to decrease as well. However, we expect the decrease to be greater moving from low to medium Fr than from low to infinite Fr.

- Kurtosis
  - We see practically the same relationships between each of St, Re, and Fr with kurtosis as we did with skew
  - As Stokes number increases, we expect the kurtosis of the spatial distribution and clustering of particles to increase.
  - As Reynolds number increases, we expect the kurtosis to decrease.
  - We also see the same relationship with Froude number, as we expect the kurtosis of the distribution to decrease when moving from low to infinite Fr, but increase when moving from medium to infinite Fr.
  - These results are also intuitively logical, since a high Reynolds number will likely cause the particles to be centered closer around the mean, thus the distribution would have less fat tails. The similar, odd relationship between Froude number with both skew and kurtosis could be due to the fact that it is not quite possible to interpret infinite gravitational acceleration and what the repercussions of that would be in terms of the particle clustering distribution.

Conclusion

Future Investigations

This short report briefly investigated "the most important unsolved problem of classical physics", and because of its brevity most likely did not add anything significant to field of turbulence. Yet, it did open the researchers mind into what can and should be explored in this field's cross with machine learning. It is important to find a good predictive model to ease the burden of the computationally expensive numeric solvers which output the data. However, if possible it would be helpful to see the relationship between Froude's and Reynold's number when it is continuous like Stoke's number. It is difficult to only have three different values for Froude's number and Reynold's number, especially when one is `Inf`.

Appendix

Model Metrics Figure

Table 3: Mean Model Error

ModelType	MSE
Linear Regression	0.000248
2nd Degree Polynomial	0.000253
Interaction Effects	0.000030
Tree Based Model	0.000014
Boosted Model	0.000163

Table 4: Variance Model Error

ModelType	MSE
Linear Regression	33003.25
2nd Degree Polynomial	35219.99
Interaction Effects	9020.97
Tree Based Model	2297.68
Boosted Model	19311.89

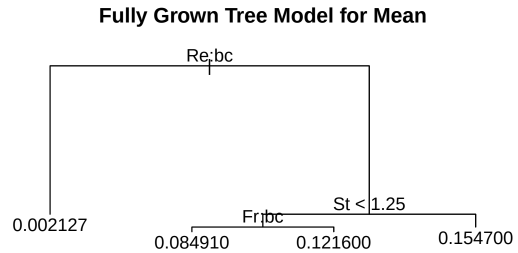
Table 5: Skew Model Error

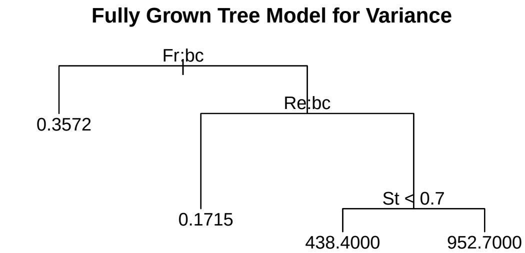
ModelType	MSE
Linear Regression	2.128693e+12
2nd Degree Polynomial	2.390292e+12
Interaction Effects	7.797398e+11
Tree Based Model	2.296011e+11
Boosted Model	1.596803e+12

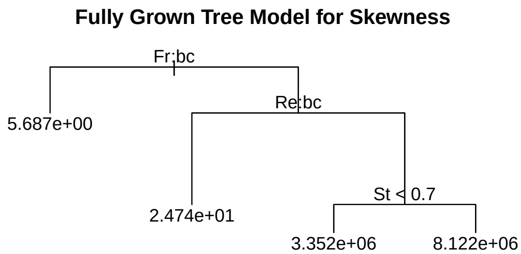
Table 6: Kurtosis Model Error

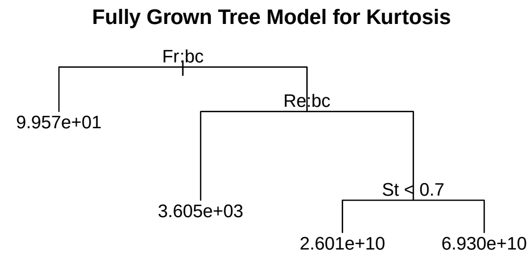
ModelType	MSE
Linear Regression	1.537524e+20
2nd Degree Polynomial	1.644865e+20
Interaction Effects	6.510998e+19
Tree Based Model	2.237499e+19
Boosted Model	1.106086e+20

Final Tree Model Plots









Simple Linear Regression Coefficients

Table 7: Mean Regression Model

	Estimate	Pr(>|t|)
(Intercept)	0.105	0.000
St	0.016	0.000
ReMedium	-0.109	0.000
ReHigh	-0.107	0.000
FrMedium	-0.010	0.052
FrInfinity	-0.013	0.006

Table 8: Variance Regression Model

	Estimate	Pr(>|t|)
(Intercept)	397.654	0.000
St	56.969	0.087
ReMedium	-231.422	0.000
ReHigh	-288.435	0.000
FrMedium	-332.974	0.000
FrInfinity	-267.527	0.000

Table 9: Skewness Regression Model

	Estimate	Pr(>|t|)
(Intercept)	3220094.7	0.000
St	528559.1	0.063
ReMedium	-1906707.5	0.000
ReHigh	-2355239.4	0.001
FrMedium	-2745078.3	0.000
FrInfinity	-2206451.9	0.000

Table 10: Kurtosis Regression Model

	Estimate	Pr(>|t|)
(Intercept)	26325470358	0.000
St	4828236326	0.046
ReMedium	-15851044054	0.000
ReHigh	-19414518255	0.001
FrMedium	-22769502054	0.000
FrInfinity	-18318122300	0.000