Task 4: Final Project

Rajan (Raj) Desai

An analysis of existing utility-scale wind & solar projects across the Continental United States.

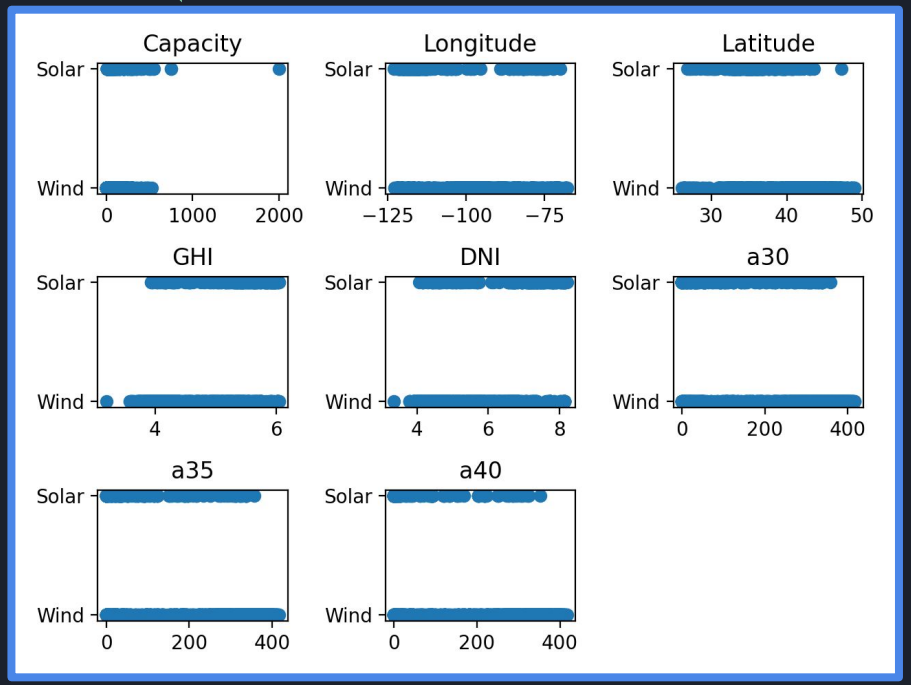
Overview

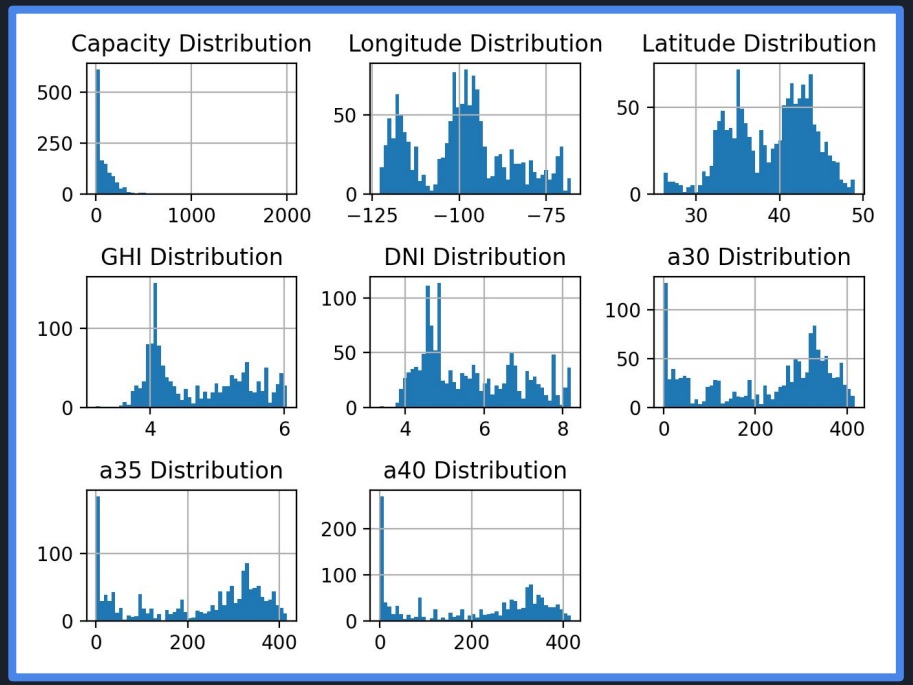
- This project seeks to analyze utility scale renewable projects (i.e. wind farms and solar farms) and seeks to answer the following questions:
- Are the current locations of these projects efficient?
    a. Are local measures of wind/solar potential good predictors for the presence of a solar vs a wind project for a given area?
- How accurate of a classifier can be built taking into account the current locations of utility scale solar and wind projects?
    a. If an accurate classifier can be built using solar/wind potential as inputs, then we can surmise that the current locations of these projects are, at the very least, sensible.
    b. If an accurate classifier cannot be built, then it is worth exploring whether federal, state, and local governments should re-evaluate their policies when building/facilitating utility scale projects in their region.

Data

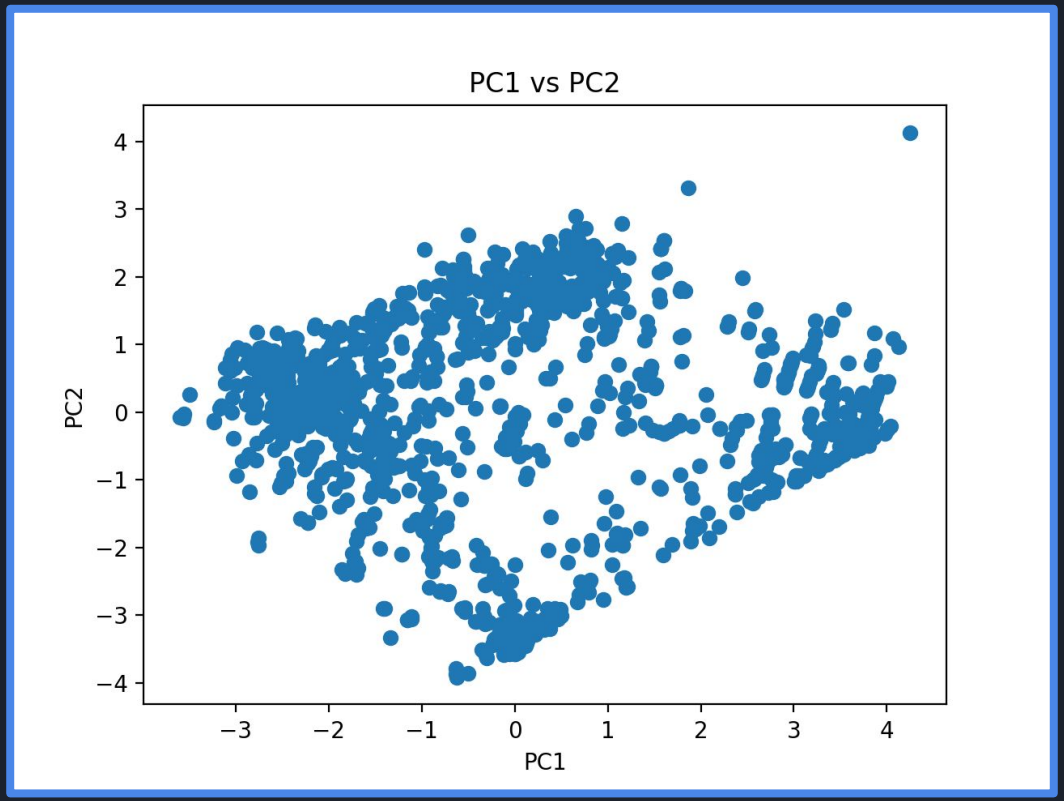
- Features:
  - Longitude/Latitude (continental U.S.)
  - Local Solar Potential
    - DNI (Direct Normal Irradiance) - Watts per square meter = Power/m^2
    - GHI (Global Horizontal Irradiance) - Watts per square meter = Power/m^2
    - Granularity: Measurements accurate to within ~240 sq miles
  - Local Wind Potential
    - Potential Installed Capacity (Megawatts)
      - a40, a35,a30
    - Granularity: Measurements accurate to within ~600 sq miles
  - Project Capacity (MegaWatts)
  - p = 8
  - n = 2009 (70% Used for training, 30% for testing)
- Supervisor
  - Type of project: 2 classes
    - Solar
    - Wind

EDA: First Impressions

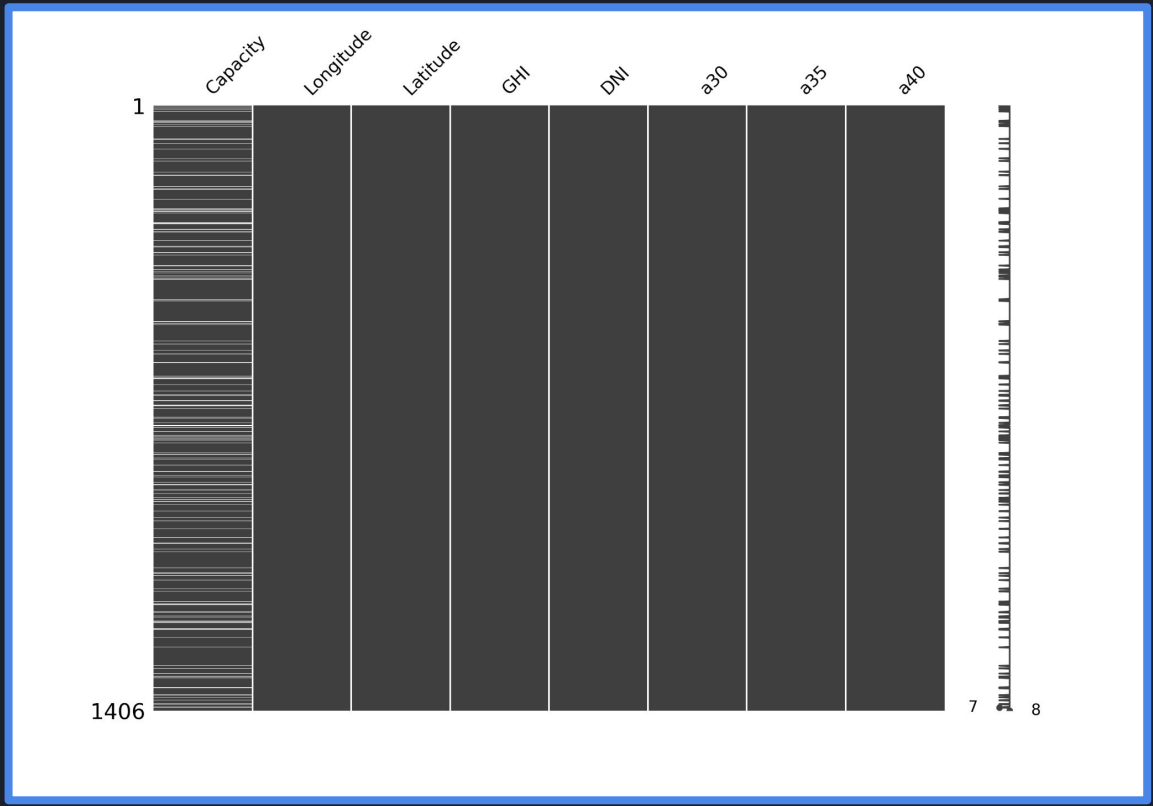




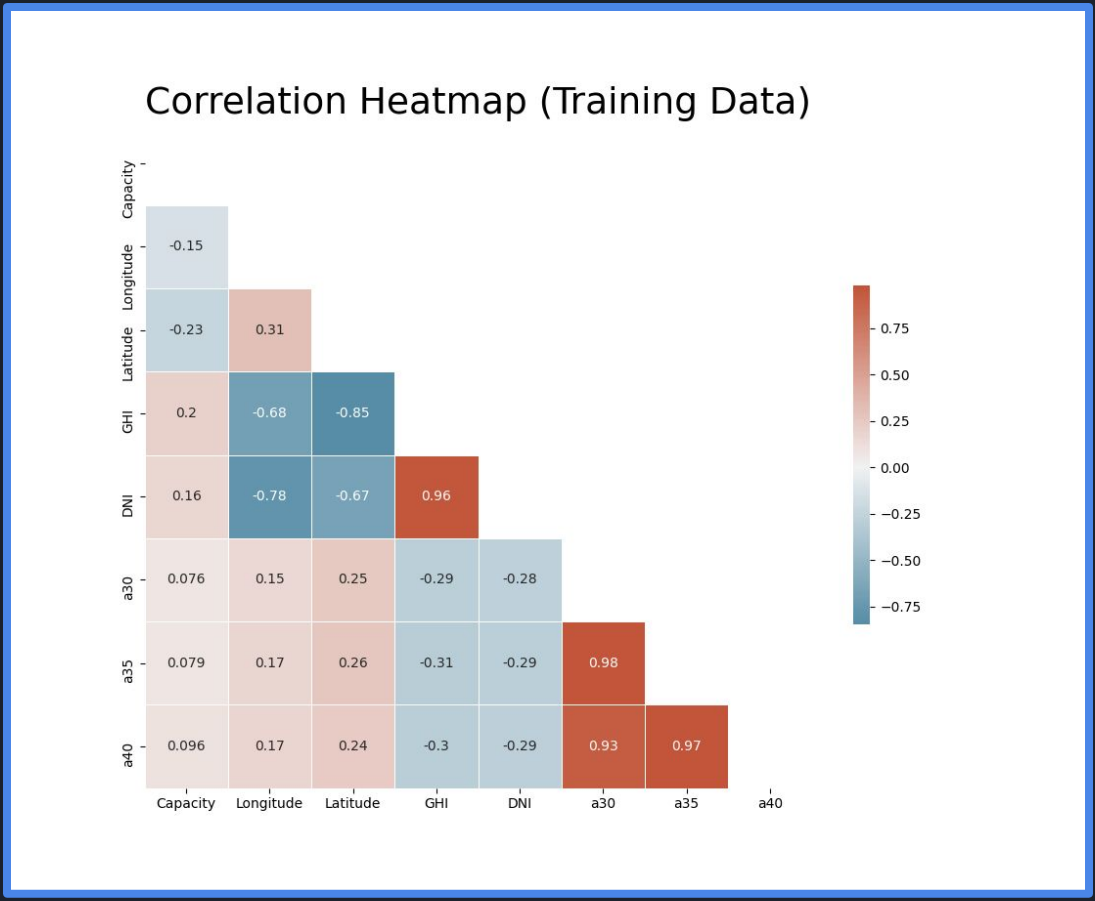
EDA: Screening For Outliers



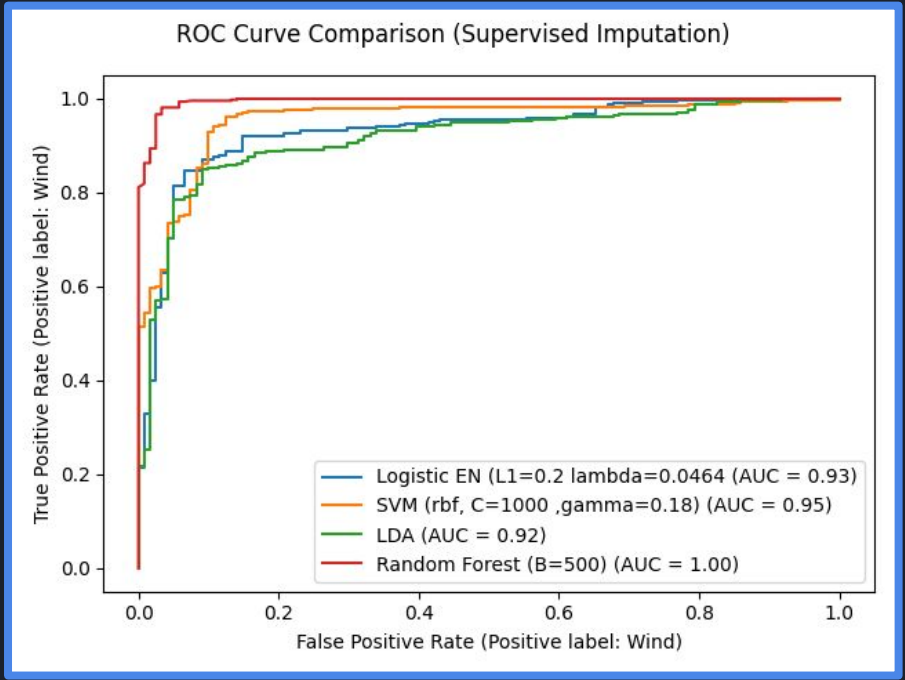
EDA: Missing Data

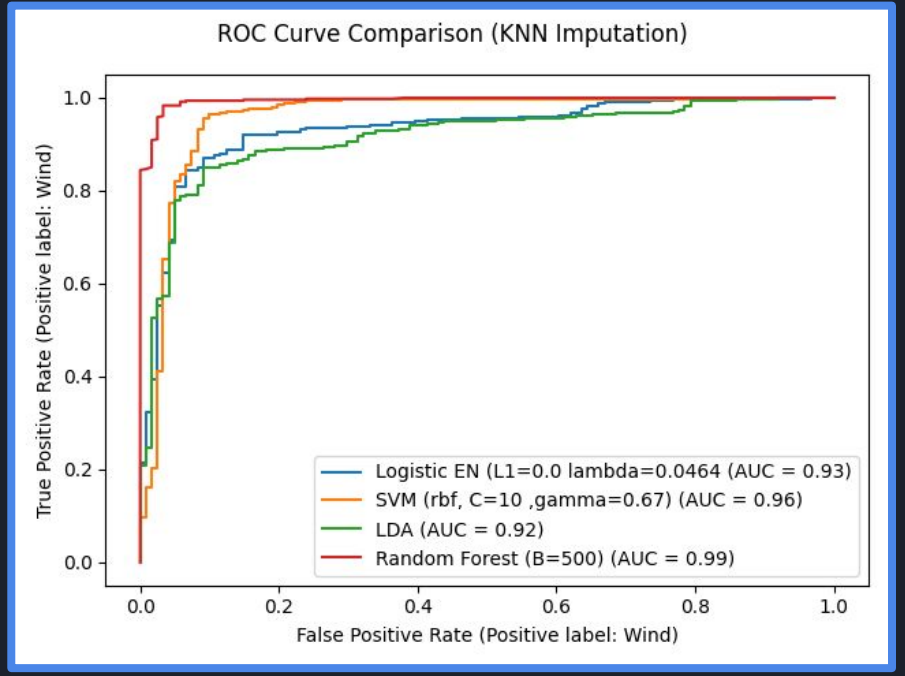


EDA: Correlation Structure

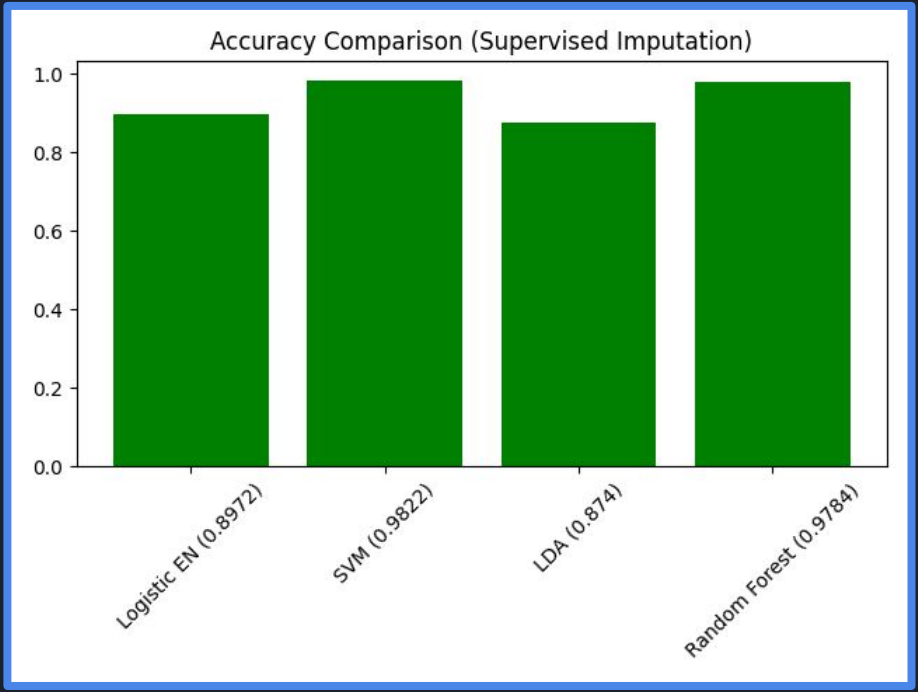


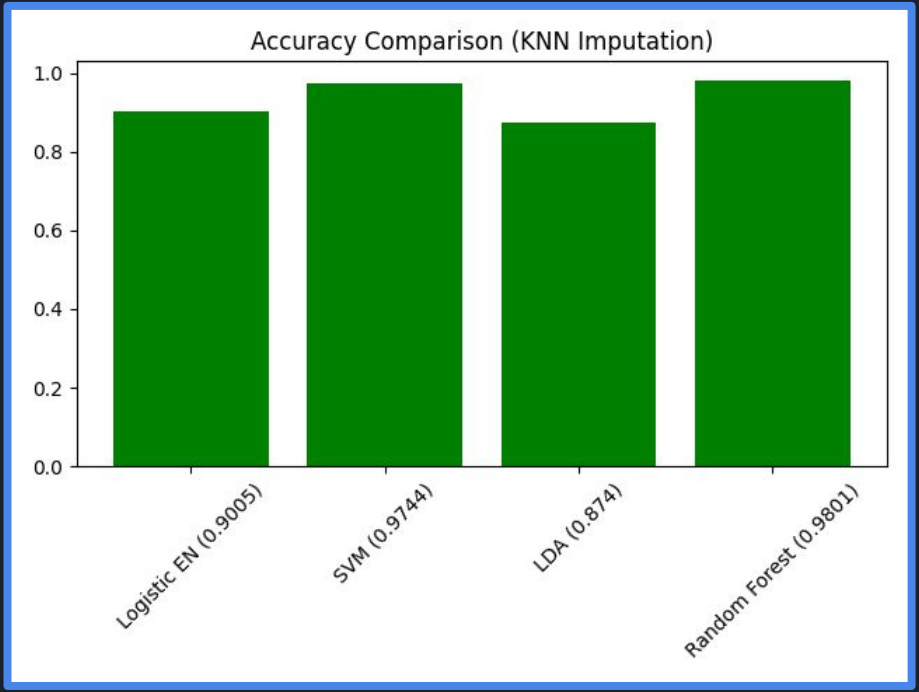
Modeling: ROC Curves



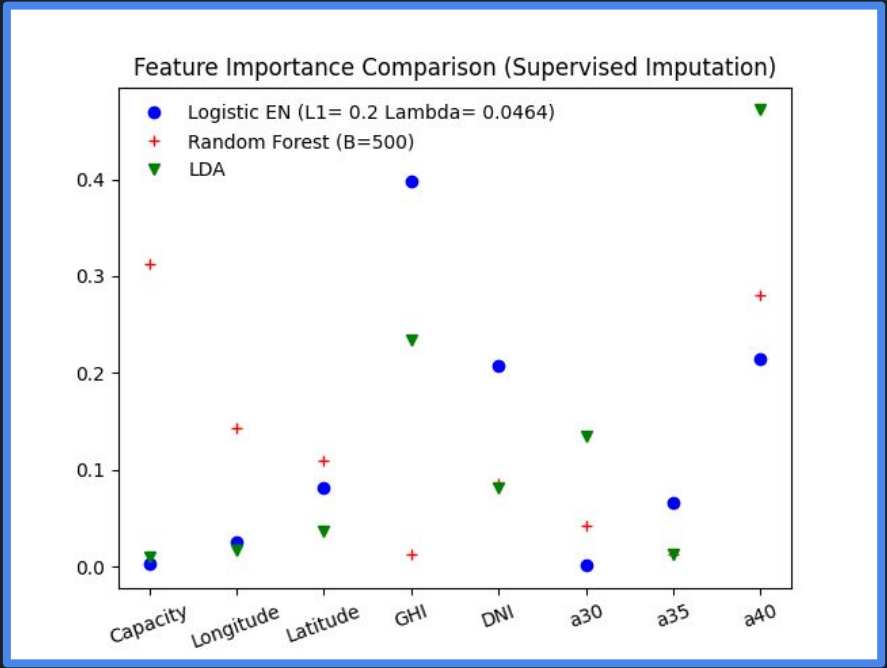


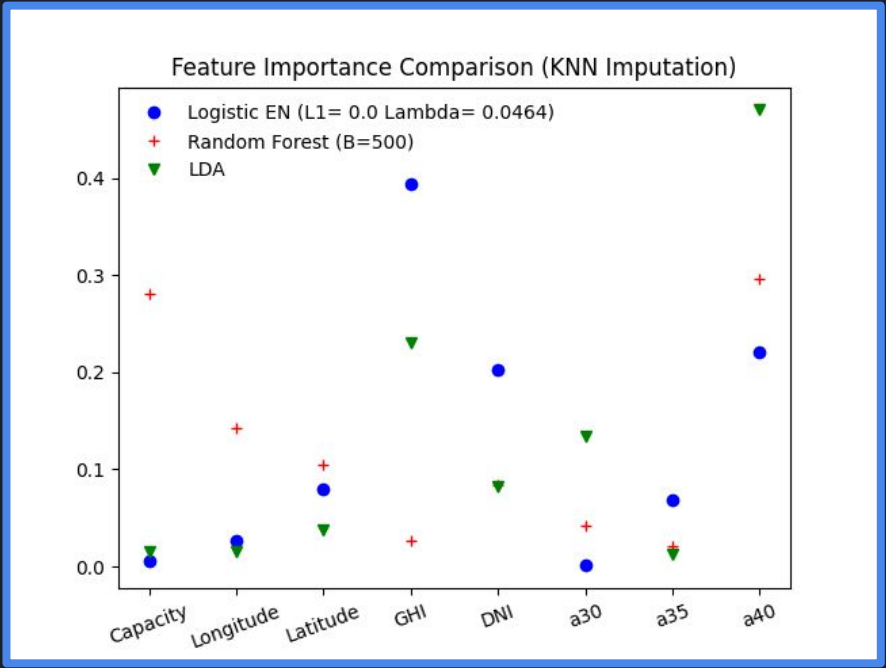
Modeling: Test Set Performance

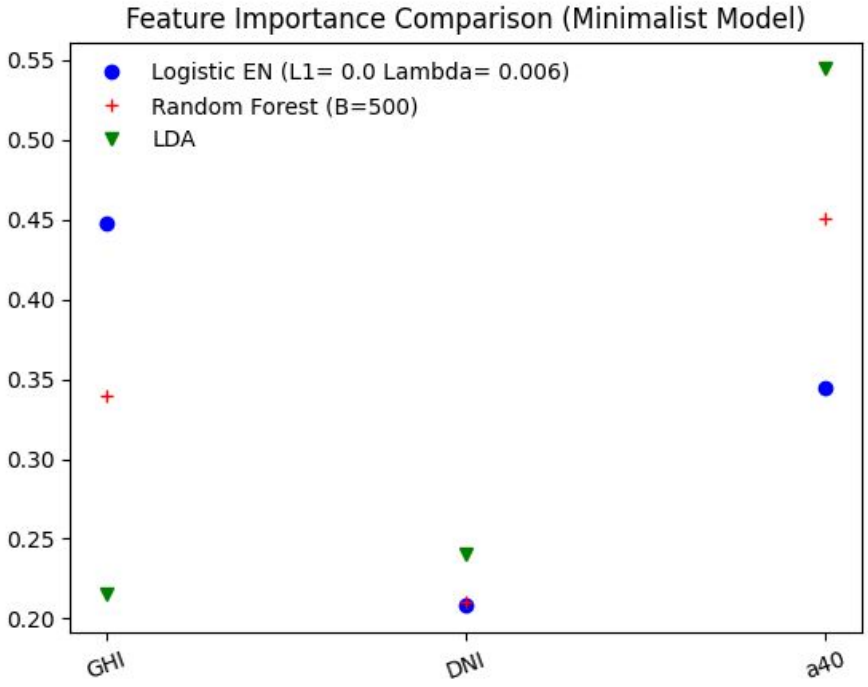


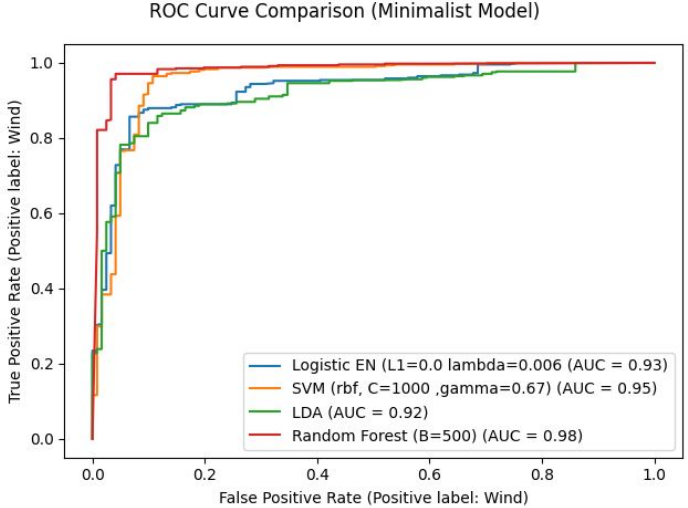


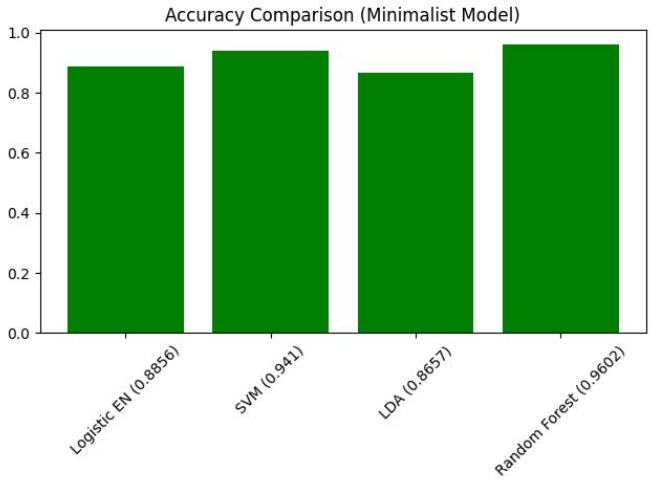
Modeling: Feature Importance

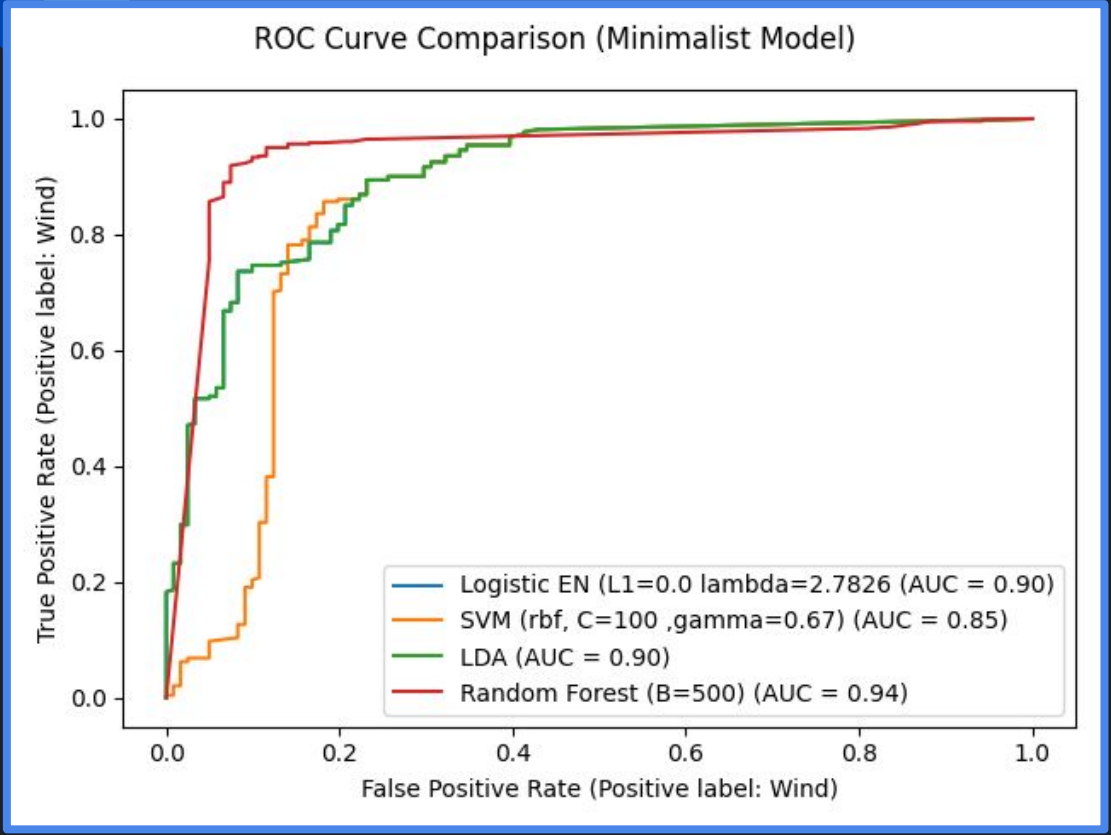


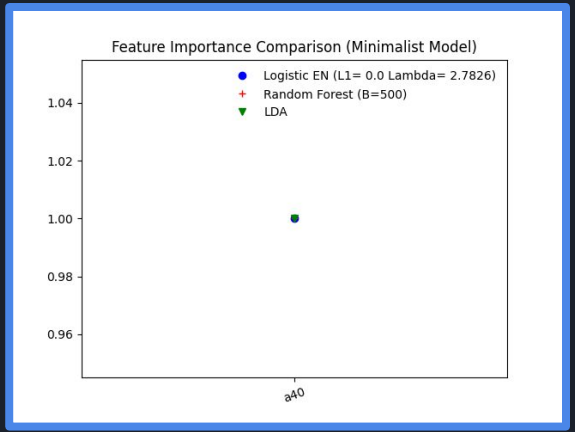


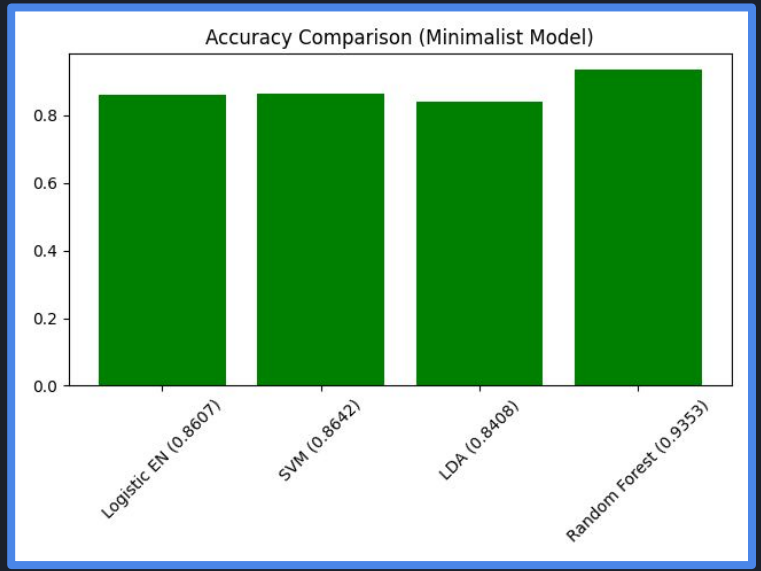

Feature Importance Comparison (Minimalist Model)
Logistic EN (L1= 0.0 Lambda= 0.006)
Random Forest (B=500)
LDA
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
GHI
DNI
a40


ROC Curve Comparison (Minimalist Model)
True Positive Rate (Positive label: Wind)
False Positive Rate (Positive label: Wind)
Logistic EN (L1=0.0 lambda=0.006 (AUC = 0.93)
SVM (rbf, C=1000 ,gamma=0.67) (AUC = 0.95)
LDA (AUC = 0.92)
Random Forest (B=500) (AUC = 0.98)


Accuracy Comparison (Minimalist Model)
Logistic EN (0.8856)
SVM (0.941)
LDA (0.8657)
Random Forest (0.9602)


ROC Curve Comparison (Minimalist Model)
True Positive Rate (Positive label: Wind)
False Positive Rate (Positive label: Wind)
0.0
0.2
0.4
0.6
0.8
1.0
Logistic EN (L1=0.0 lambda=2.7826 (AUC = 0.90)
SVM (rbf, C=100 ,gamma=0.67) (AUC = 0.85)
LDA (AUC = 0.90)
Random Forest (B=500) (AUC = 0.94)


Feature Importance Comparison (Minimalist Model)
Logistic EN (L1= 0.0 Lambda= 2.7826)
Random Forest (B=500)
LDA
1.04
1.02
1.00
0.98
0.96
a40


Accuracy Comparison (Minimalist Model)
0.0
0.2
0.4
0.6
0.8
Logistic EN (0.8607)
SVM (0.8642)
LDA (0.8408)
Random Forest (0.9353)

Lessons Learned

1. Plan better
    a. The initial idea of the project was nice, but I should have made sure that all of the data existed in an accessible format prior to starting
    b. Ended up facing a lot of file format hurdles, as well as compromising on which features to use.
2. Data detail
    a. Data can go through a LOT of manipulation before entering a model, even for a small project like this.
    b. I didn't realize that the granularity of my wind/solar potential was that big until just yesterday.
3. Utility scale projects do appear to be in efficient locations
    a. Some limitations due to class imbalance, but measures of wind potential (a40) and solar potential (DNI, GHI) are some of the best predictors

Biggest Challenge

- Getting the dataset together:
  - Solar Farms names and addresses were scraped from a website
    - Queried a lat/long coordinate for each farm
  - Match up solar/wind potential numbers with solar and wind farms
  - Lot of data cleaning to make sure all values were in the same format
- 80-90% of time on this project was spent:
  - Learning the necessary Python code
  - Collecting and cleaning the data

If I had to do it over...

- Spend more time ironing out my scientific question
- Plan better
  - Verify all data and data sources before beginning
- Look into measures to account for class imbalance

Thank You So Much!!!

636 and 656 were easily in my top 3 favorite classes in this program :)