Covid 19 Death Prediction

94879 OAI
Final Presentation

Bingjie (Grace) Liu
Yiwen (Cathy) Cheng
Haodong Zhang
Rugved Somwanshi

Initial Analysis of Data and Decisions

- Lots of null values (Bivalent vaccine NA columns dropped, Booster vaccine data NA filled with 0, other rows with NA were dropped)

- Death count was highly proportional to the population size of the county, leads to skewed data where population was higher, deaths were higher.

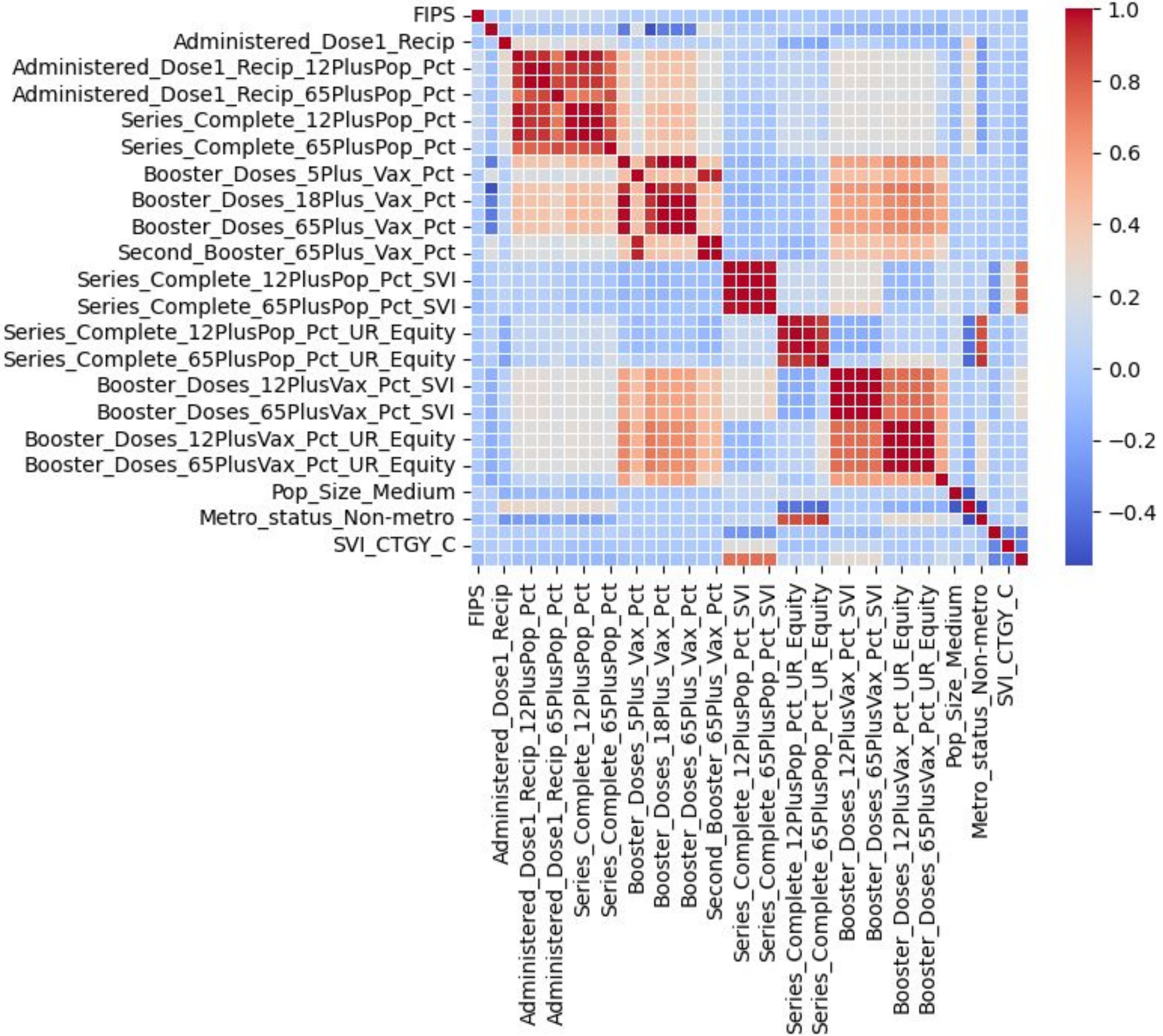
- Used Death % as the target column to normalize the population size to help us predict the effect of vaccination on deaths

Feature Engineering

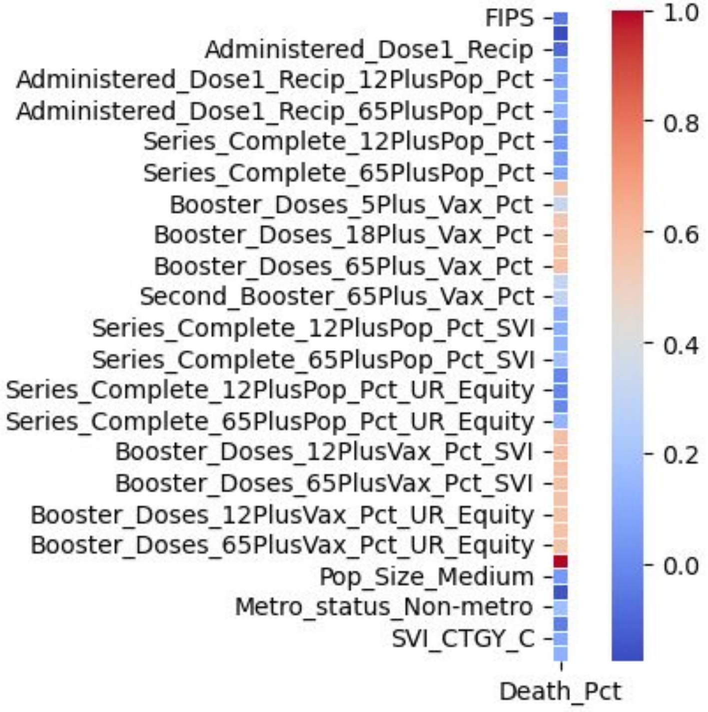
- Only kept columns that are related to percentage data, e.g. Percentage of 18+ population with a complete series.
- Actual numbers are highly correlated with population sizes -> could impact model prediction accuracy.

```
# Calculate the 33rd and 66th percentiles
percentile_33 = joined_df['Census2019'].quantile(0.33)
percentile_66 = joined_df['Census2019'].quantile(0.66)

# Define the bins and labels
bins = [0, percentile_33, percentile_66, joined_df['Census2019'].max()]
labels = ['Small', 'Medium', 'Large']

# Categorize the population based on its size
joined_df['Pop_Size'] = pd.cut(joined_df['Census2019'], bins=bins, labels=labels, include_lowest=True)
dummies = pd.get_dummies(joined_df[['Pop_Size', 'Metro_status', 'SVI_CTGY']], drop_first=True)
joined_df = pd.concat([joined_df, dummies], axis=1)
```

Overall Feature Correlation



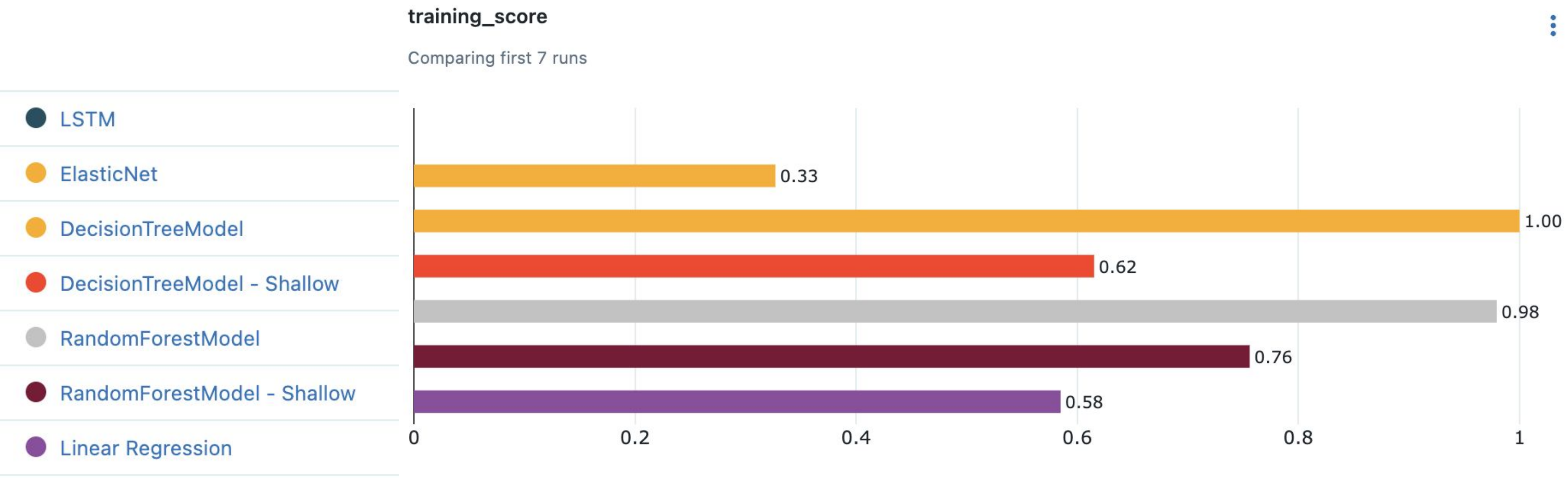
How are features related to Death %?



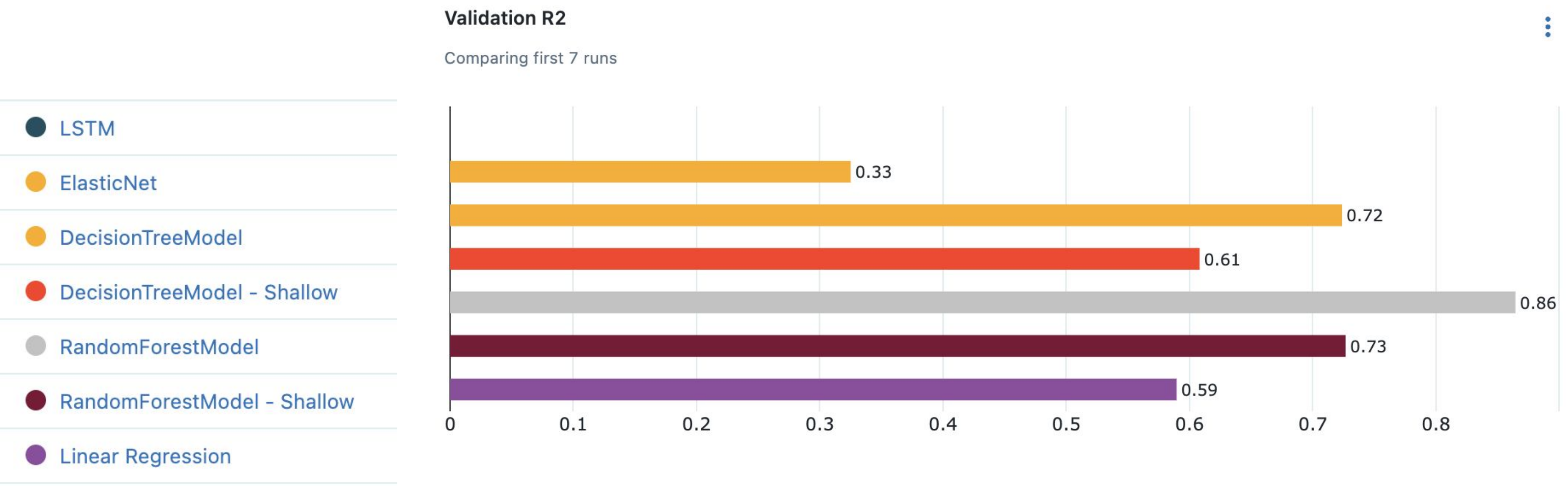
Models we chose to train

1. Linear Regression
2. Decision tree
3. Decision tree with less depth and number of leaves
4. Random Forest
5. Random forest with less depth and number of leaves
6. LSTM

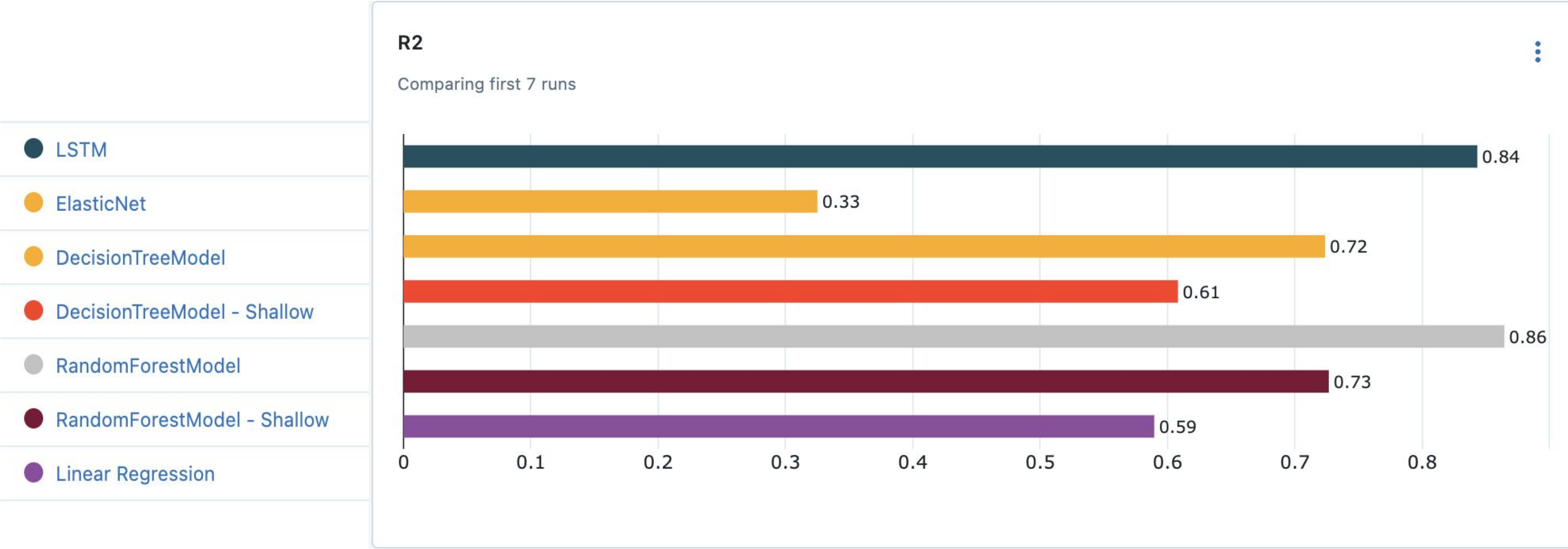
Comparing Training R2 Score



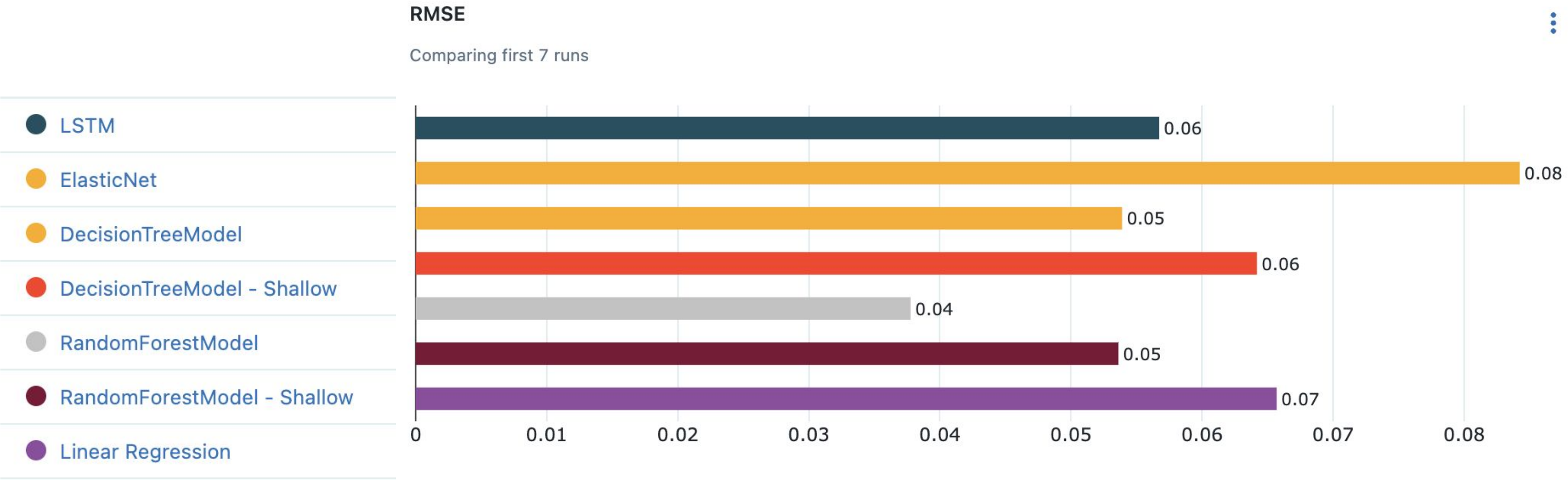
Comparing Validation R2 Score



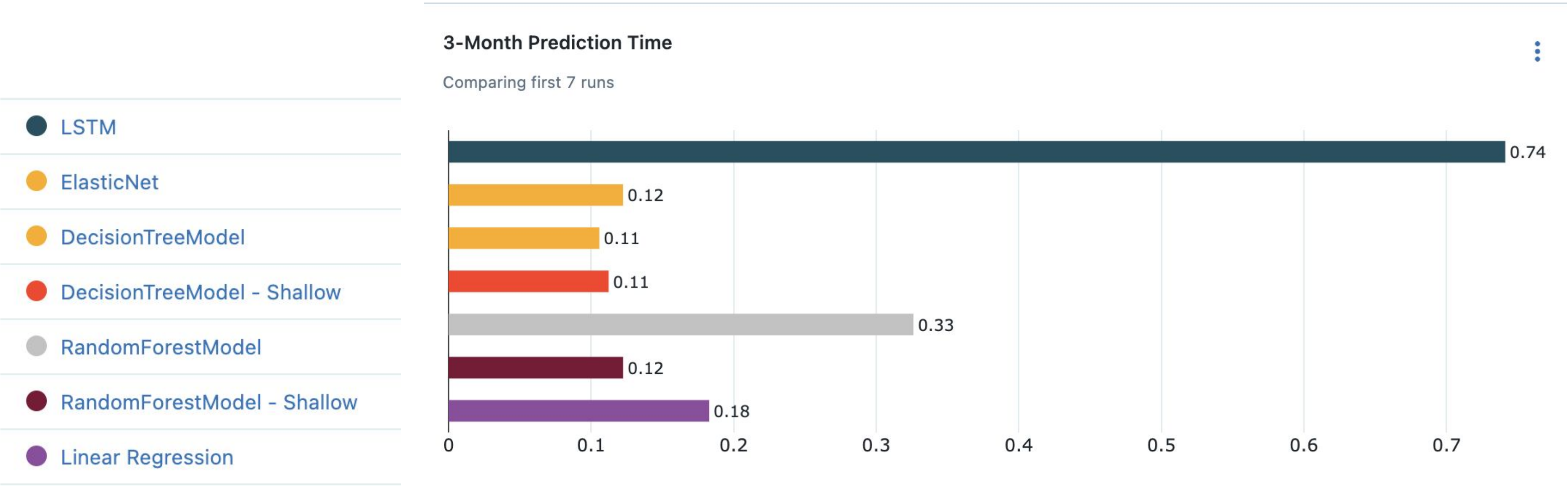
Comparing R2 Scores



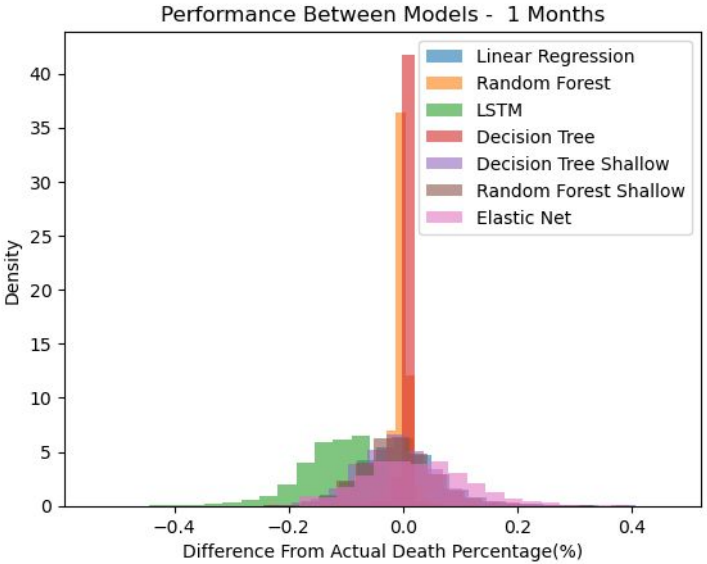
Comparing RMSE

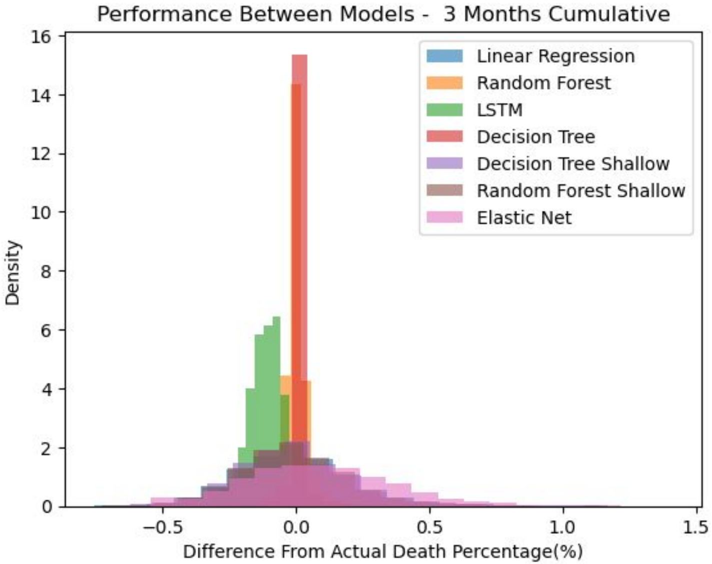


Comparing Prediction Time



Distribution of Actual - Predicted Differences



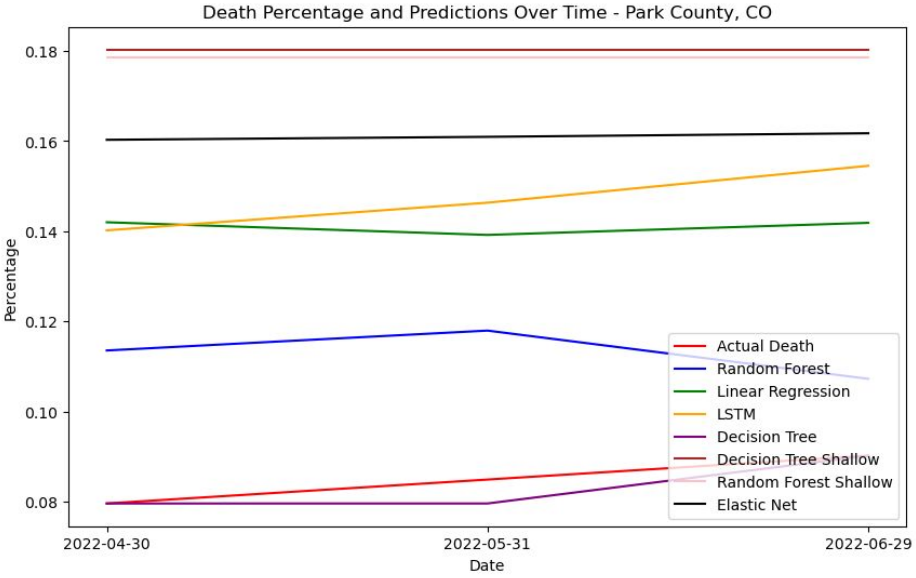


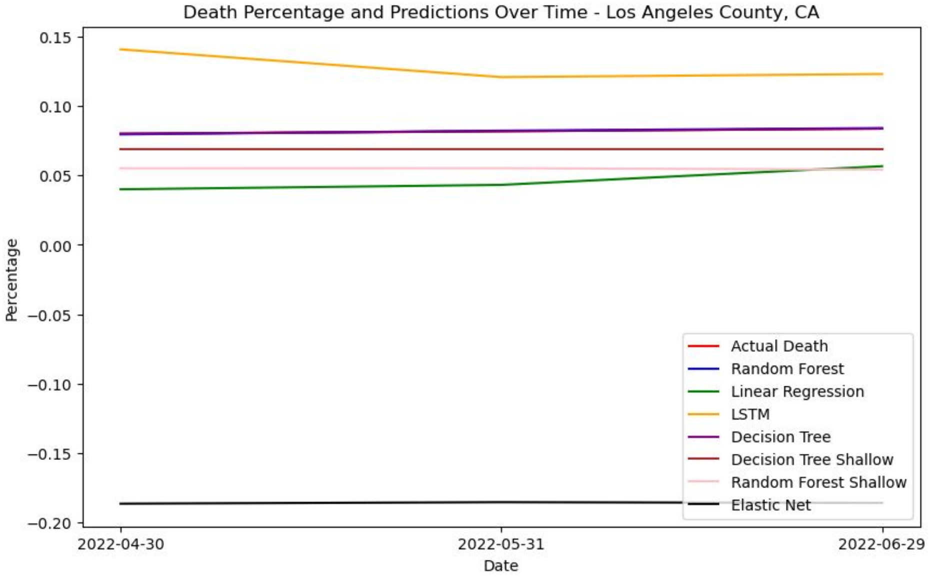
- Model fits better if its histogram is roughly normal and has a significant high peak close to 0.
- Decision Tree model's predictions were closest to the actual values more frequently than the other models, However, it is overfitted (training score = 1).
- Random Forest Model performing better than all other models.
- LSTM is constantly underestimating for all time period.

Prediction V.S. Actual for Sample Counties

Park County, CO
2019 Population: 18,845
Apr, May, Jun Death: 15, 16, 17

Los Angeles County, CA
2019 Population: 10,039,107
Apr, May, Jun Death: 8040, 8214, 8406





- Random Forest model has the best prediction (besides the overfitted decision tree)
- Random Forest is more accurate in predicting deaths for large counties (One possible reason is that individual effect is less significant in large counties)

Moving Forward

1. If we use the concept of Challenger-Champion models, Random Forest should be deployed after getting hyper-parameter tuning done.
2. But we must not let go of models like the overfitted Decision Tree and under fitted LSTM.
3. Both of these other models will improve overtime as we get more data
4. They can even surpass the accuracy of the Random forest model and could be made the new champion models

Thank you!