Data Science Project phase – 3

Nikitha Narsing

1. **In Texas state cases,**

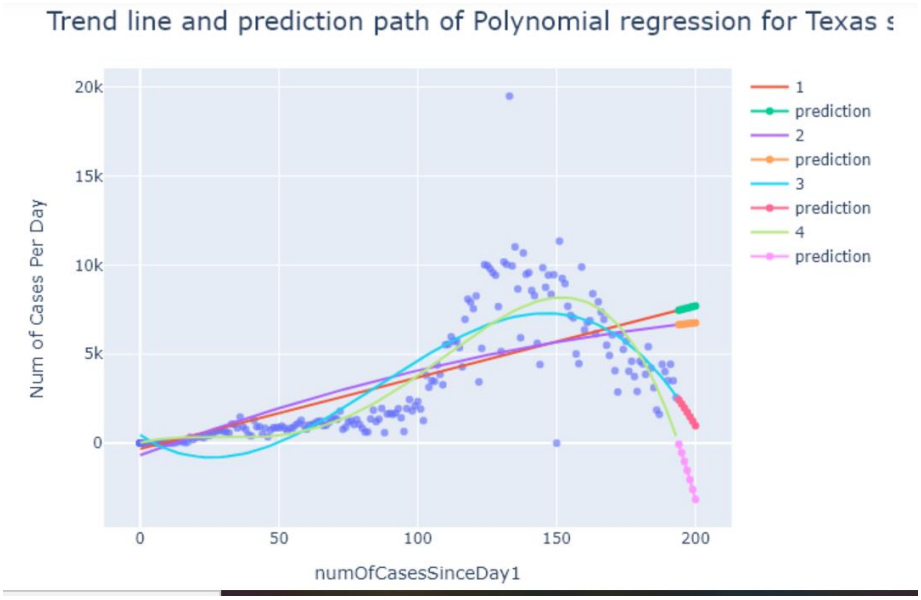
Linear regression stats:

```
print("Mean Absolute Error:",metrics.mean_absolute_error(y,cases_prediction))
print("Mean Squared Error:", metrics.mean_squared_error(y,cases_prediction))
print("Sqrt of Mean Squared Error:",np.sqrt(metrics.mean_squared_error(y,cases_prediction)))
print("R^2 Score:",r2_score(y,cases_prediction))
```

```
Mean Absolute Error: 1487.979538079421
Mean Squared Error: 5086892.792805633
Sqrt of Mean Squared Error: 2255.414106723116
R^2 Score: 0.5553553490553019
```

Polynomial regression with degree 3 is the best fit for the data with the below stats.

```
R-square of state cases for degree 3 : 0.739881967988645
RMSE of state deaths for degree 3 : 1725.063102713248
```



2. **In Texas state deaths,**

Linear regression stats:

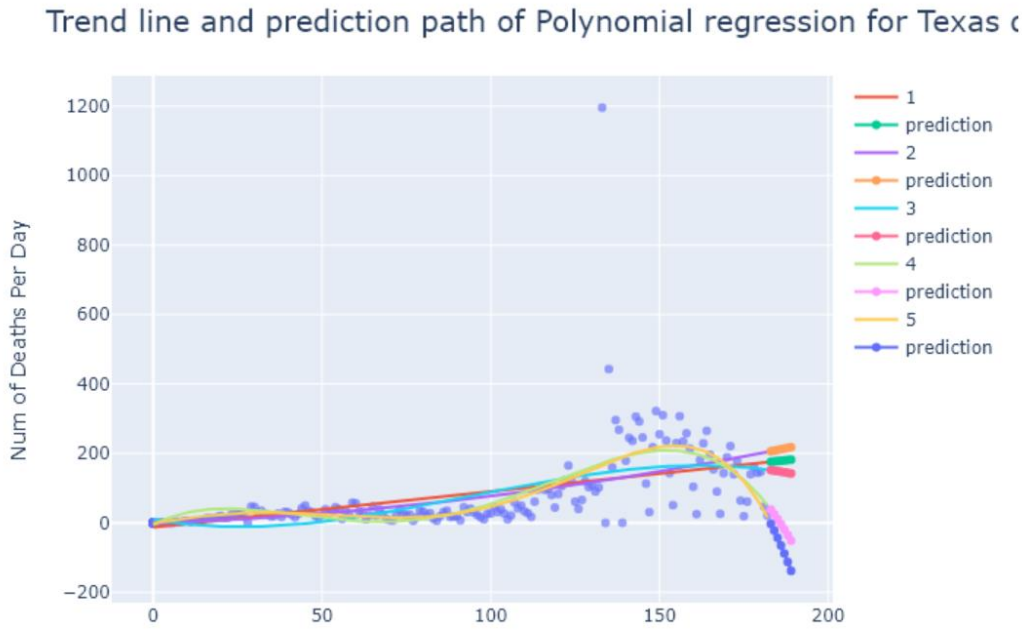
```
print("MSE score:",metrics.mean_squared_error(y_d,cases_prediction))
print("MAE score:",metrics.mean_absolute_error(y_d,cases_prediction))
print("RMSE score:",np.sqrt(metrics.mean_squared_error(y_d,cases_prediction)))
print("R^2 score:",r2_score(y_d,deaths_prediction))
```

```
MSE score: 13892767.415642764
MAE score: 2927.841239089664
RMSE score: 3727.300285145103
R^2 score: 0.31807891821027645
```

Polynomial regression stats with degree 3 and 4:

```
R-square of state deaths: 3 0.3561712773689837
RMSE of state deaths for degree 4350.863509042784
R-square of state deaths: 4 0.44046784900021896
RMSE of state deaths for degree 4346.846638264234
```

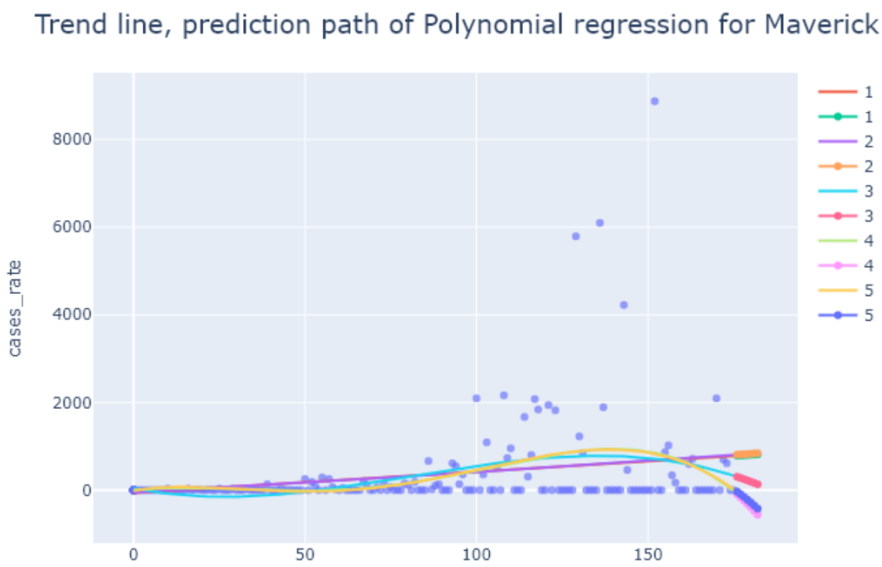
Though the R-square value seems to be higher the plot seems to be overfitting for degree 4 regression. So, the best fit is degree 3.



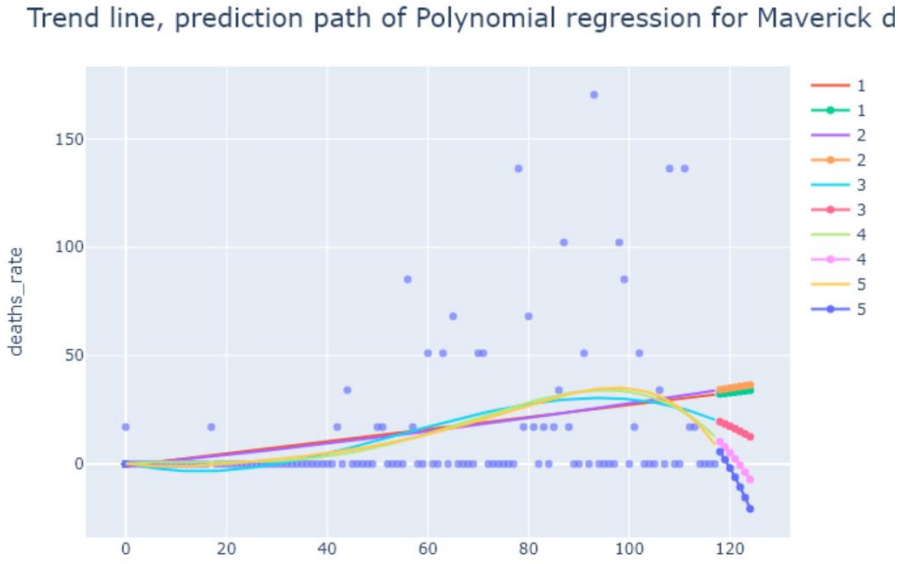
2. Identify which counties are most at risk.

As per my poly regression model with degree 3, maverick county is more at risk when it comes to cases and deaths as the predicted deaths are higher than other counties.

Cases: Predicted cases are between 0 and 1000 whereas other counties predicted deaths are between 0-500



Deaths: Predicted Deaths are between 0 and 50 whereas other counties predicted deaths are between 0-20



3. **According to my prediction among 5 states using polynomial regression with degree 3 on state deaths, the graph looks well fitted on degree 3 for all states.**

   But no state until next month each day would reach its point of no return as the ICU beds are way more higher than the predicted deaths on each day.

4. **Compare RMSE of Decision Tree, Random Forest and ARIMA.**

```
rmse=np.sqrt(mse)
print("RMSE score of Decision tree", rmse)
```

```
RMSE score of Decision tree 1845.9943507110586
```

```
rmse=np.sqrt(mse)
print("RMSE of Random forest regressor: ", rmse)
```

```
RMSE of Random forest regressor:  1374.4617767061195
```

```
error = metrics.mean_squared_error(test, fc_series)
rmse=np.sqrt(error)
print('Test MSE: %.3f' % error)
print('RMSE of ARIMA: %.3f' % rmse)
```

```
Test MSE: 895523.708
RMSE of ARIMA: 946.321
```

We can see that RMSE value of ARIMA < Random Forest Regressor < Decision tree, so ARIMA is the best model for this timeseries data.

5. Use 5 different variables from the enrichment data to predict the spread rate (cases and deaths) of COVID-19 in a county. Compare Random Forest and Decision Trees (RMSE error).

Decision Tree:

Decision tree RMSE score

```
mse = metrics.mean_squared_error(y_test,y_pred)
rmse=np.sqrt(mse)
print("Decision tree RMSE:" ,rmse)
```

```
Decision tree RMSE: 2081.5717523064154
```

Feature importance of Decision tree

```
importance = regress.feature_importances_
print("Decision Tree feature importance:",importance)
```

```
Decision Tree feature importance: [0.11728634 0.86300527 0.01091517 0.00879322]
```

Random Forest:

```
mse = metrics.mean_squared_error(y_test,prediction)
rmse=np.sqrt(mse)
rmse
```

552]: 1878.6287301603688

Feature importance of RF

```
importance = rf.feature_importances_
importance
```

553]: array([0.23371195, 0.56293294, 0.17145357, 0.03190154])

As we can see the variable March employment has the highest relative importance among all the variables for both decision tree and random forest regressors, so if people could work from home there would be less spread of the virus as employment affects the cases in a county.