

Predicting the Accident Severity from Sample Dataset

Shalesh Nath Sharma

August 22, 2020





Business Problem

- In this project we will predict accident severity based on different features. This report will be targeted to Police officer to make them more wary for these particular places.

Data

- In this project we will use sample dataset "Data-Collision" to solve the problem.

Data Cleaning

We will drop these columns :

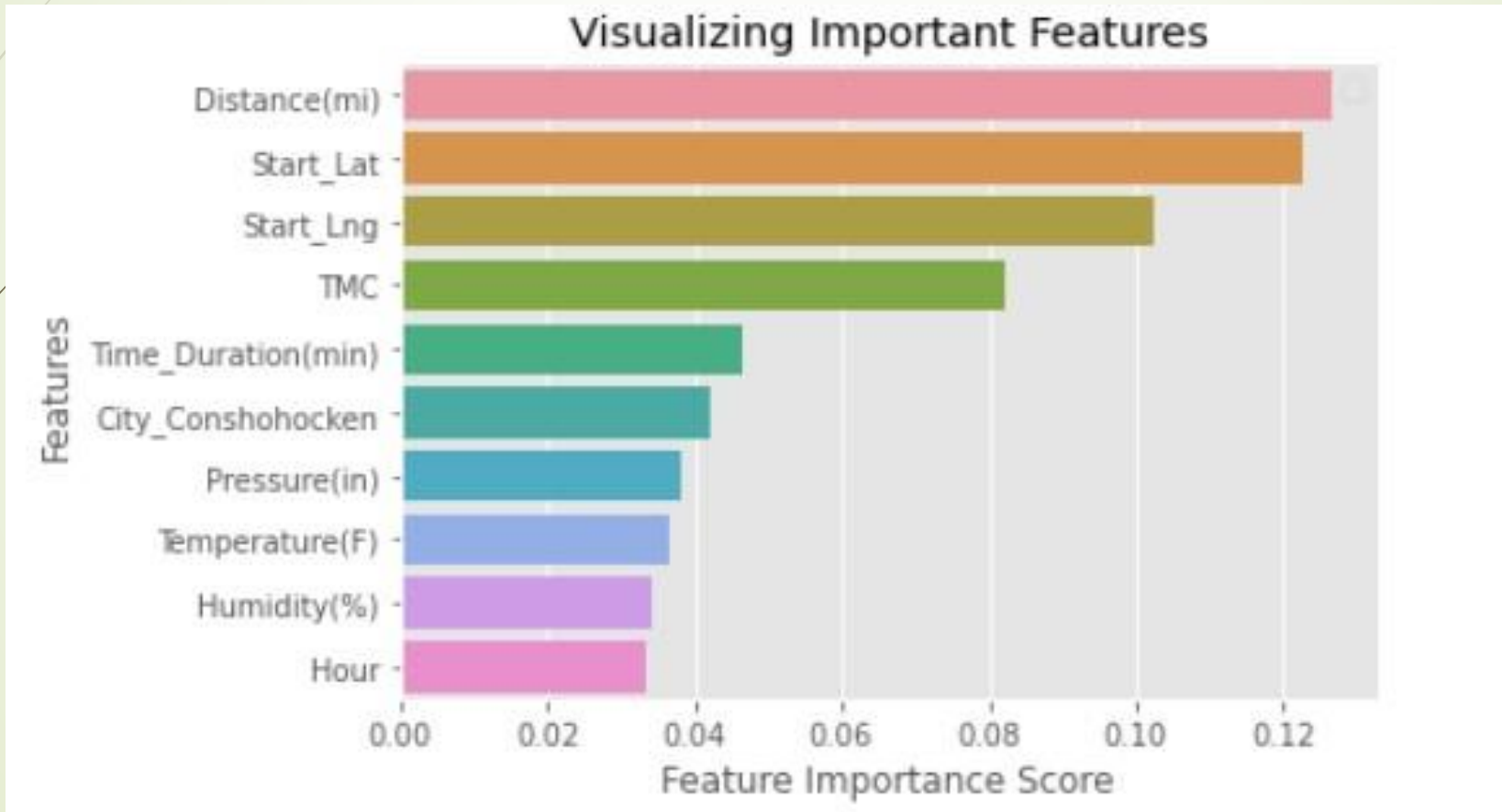
- INCKEY, COLDETKEY, STATUS : I don't know the importance of matching unique key and secondary key, would be appreciated if someone could tell me as I can't find it on internet.
- OBJECT ID, REPORTNO, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOTCOLNUM : Not relevant to predict Severity
- SEVERITYCODE.1 : Duplicate of SEVERITYCODE
- SEVERITYDESC, SDOT_COLDESC, ST_COLDESC: Description from another columns
- INCDATE, INCDTM : For simplicity
- SEGLANEKEY, CROSSWALKKEY : Too many unique Value



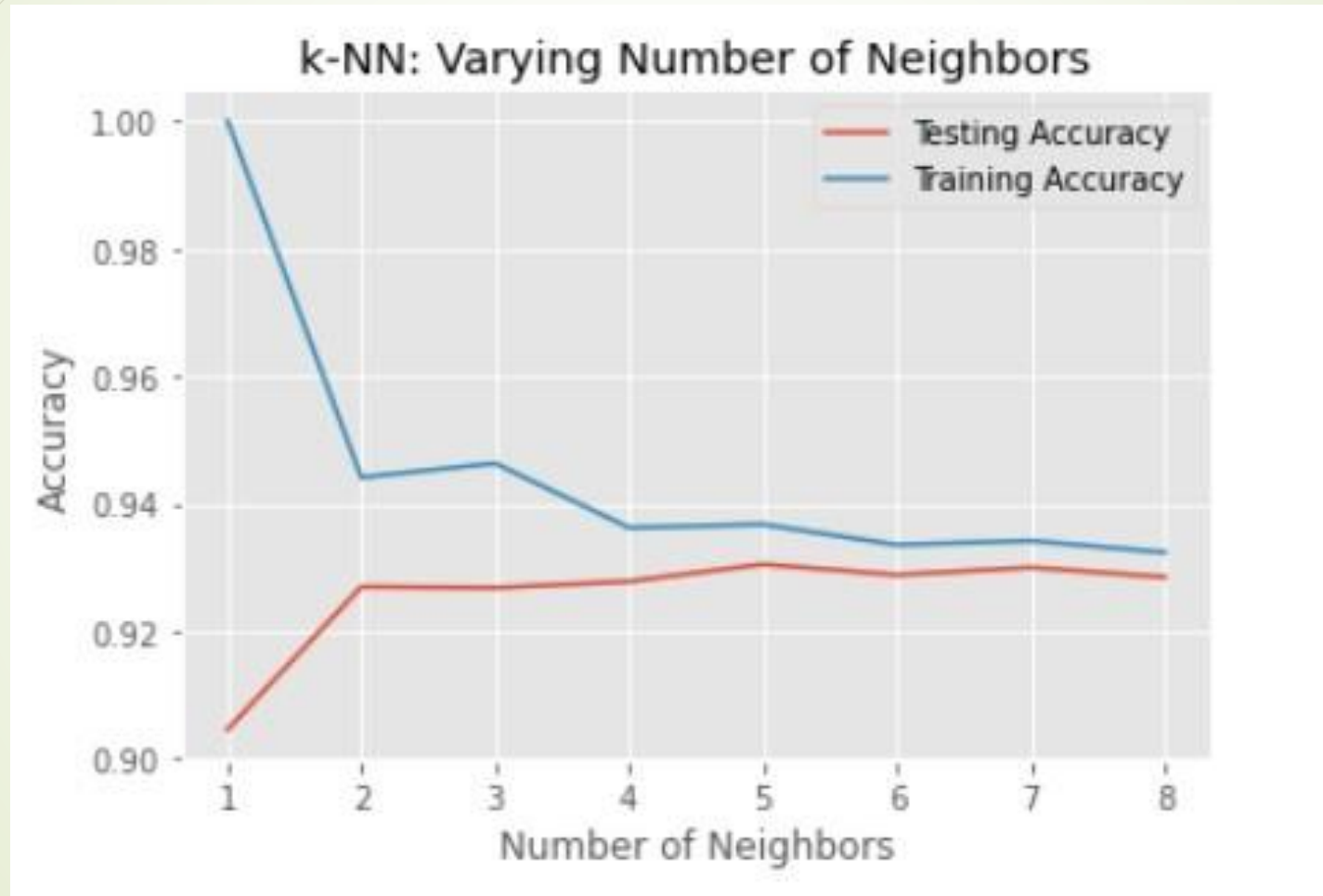
Feature Selection

- We assume Location where collision happen will be the most important feature to predict accident severity. Followed by Road Condition, Light Condition, Weather and Description of collision. Based on our assumption We will focusing on location (Junction type, Collision Type)

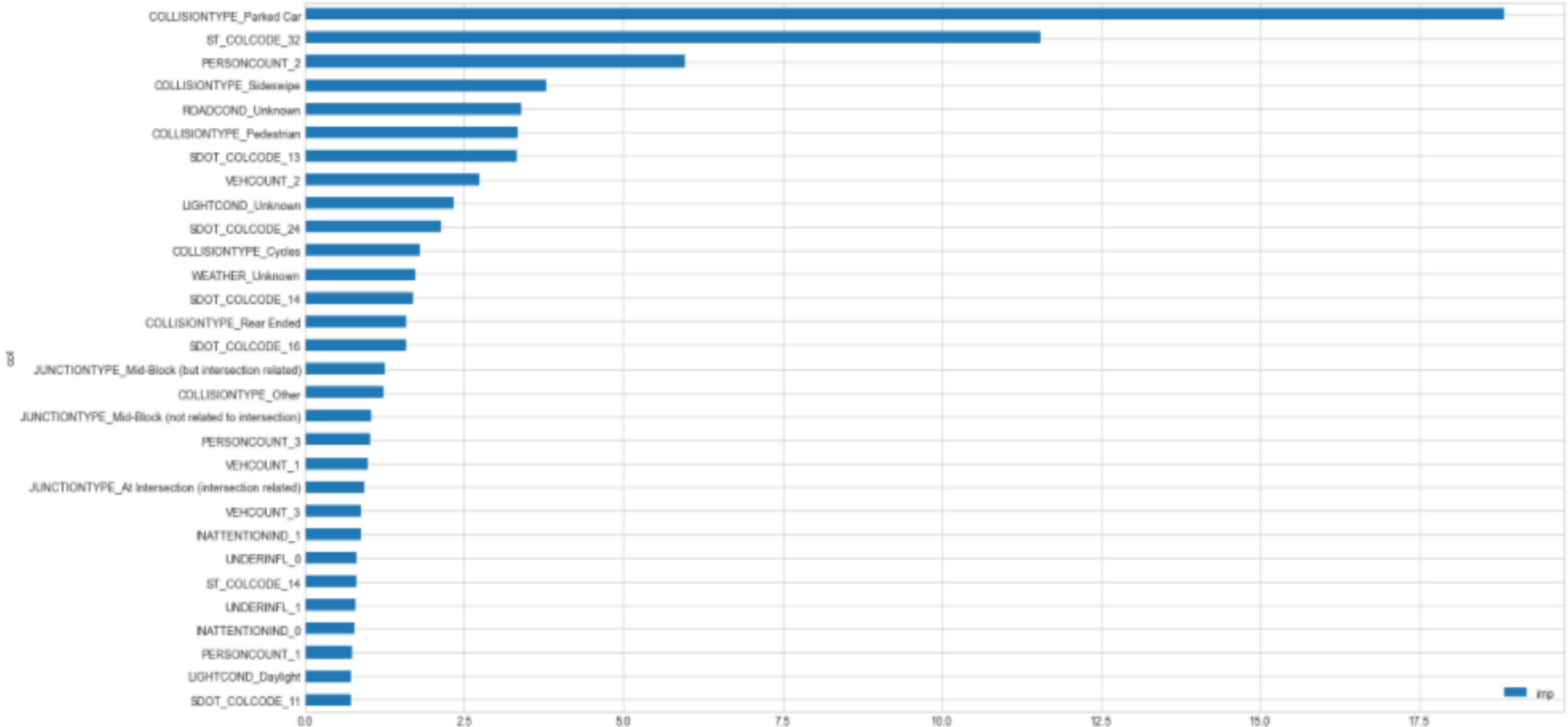
Target Feature : Severity



KNN Neighbours



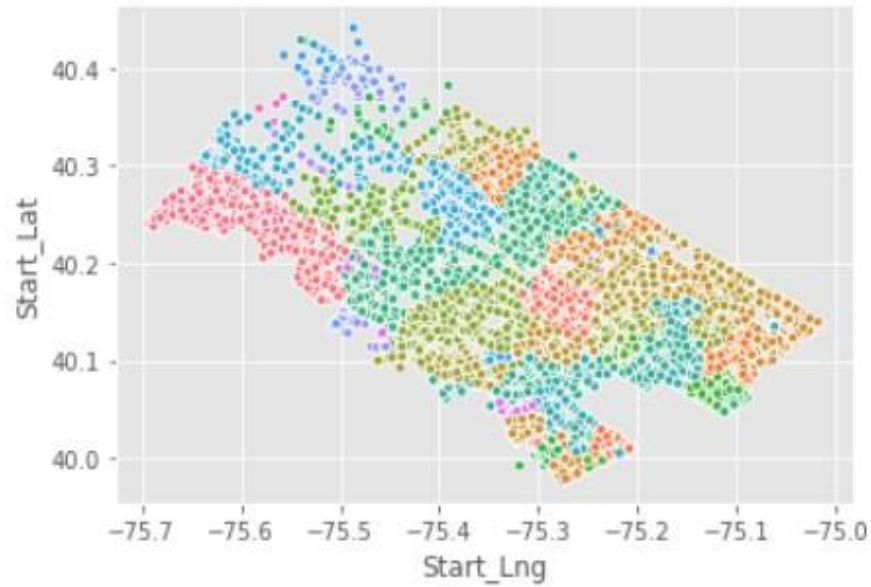
Most Important features in the dataset



Colour wise clusters by city

In [24]: *# Map of accidents, color code by city*

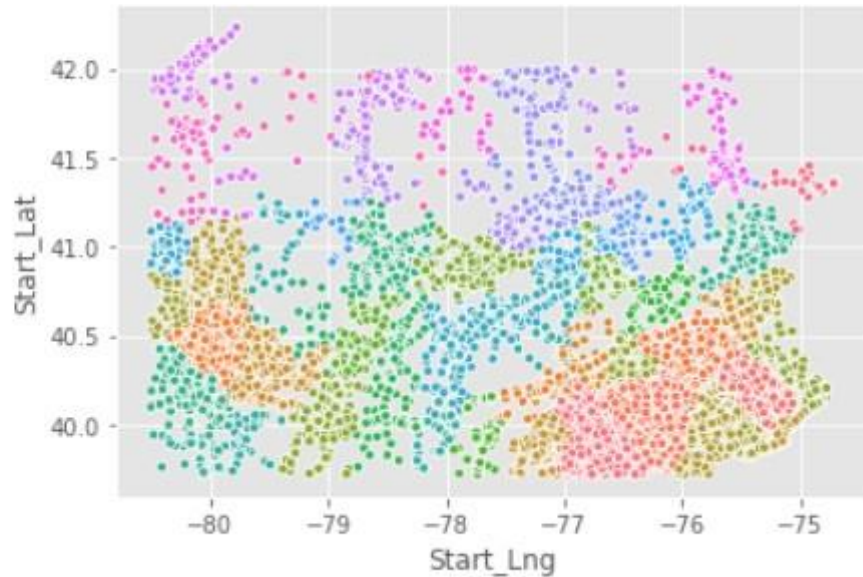
```
sns.scatterplot(x='Start_Lng', y='Start_Lat', data=df_county, hue='City', legend=False, s=20)  
plt.show()
```



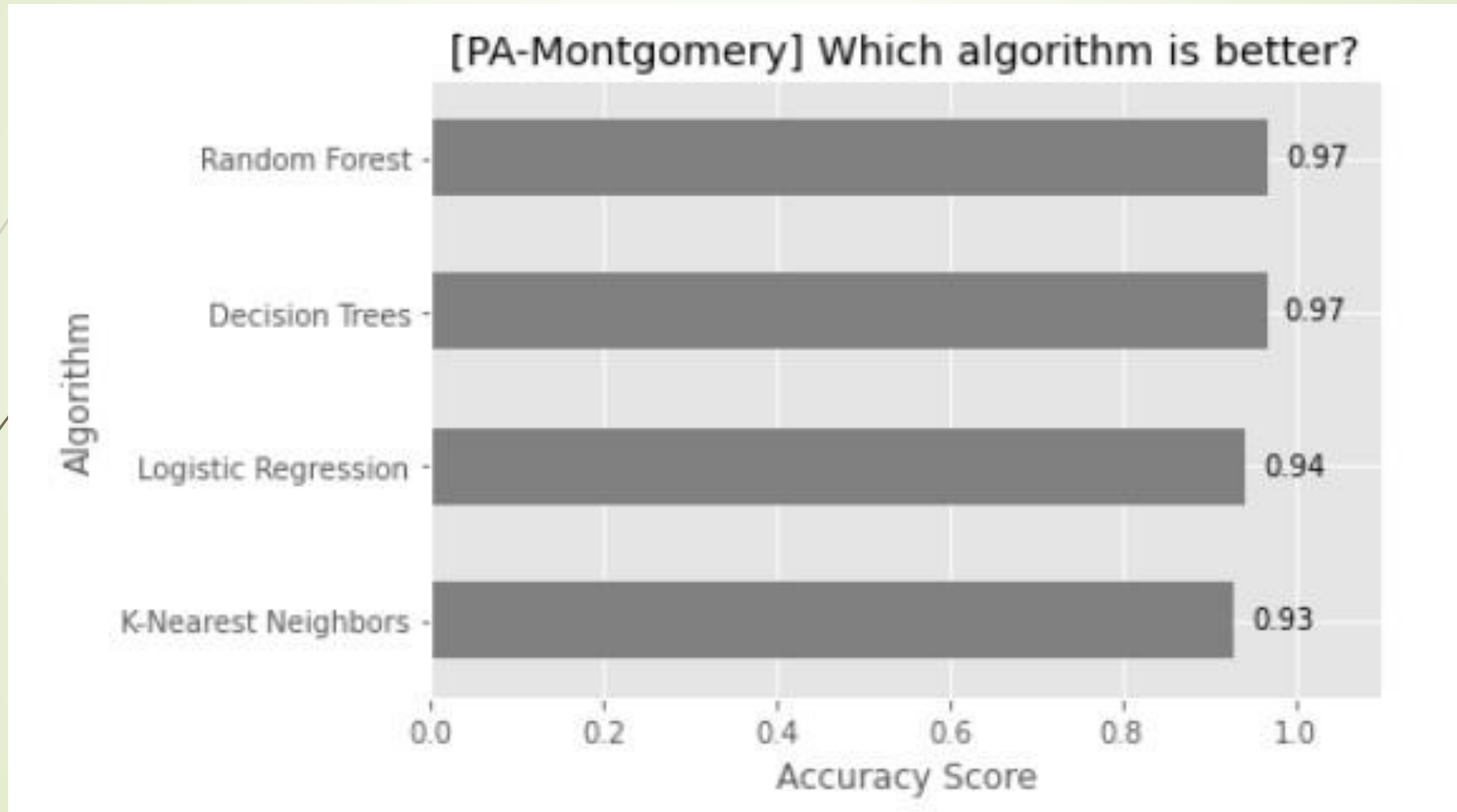
Colour wise clusters by country

In [22]: *# Map of accidents, color code by county*

```
sns.scatterplot(x='Start_Lng', y='Start_Lat', data=df_state, hue='County', legend=False, s=20)  
plt.show()
```



Final Result from different predictive models






Result and Discussion

- ▶ our ML model manage to predict severity of car accident with 97% accuracy, with the most important feature is Collision type with parked car this is make sense because when there is accident with another parked car most likely it will be a property damage collision.
- ▶ some feature that I dont use such as INCDATE,INCDTTM could be used to predict whether the accident more likely happen at night, day, weekend, weekdays, etc. We could also group up Person Count when the number reach more than 5, and if the target feature has more than 2 variable it probably more useful in real life (such as 3 = fatality, 2 = serious injury, 1 = injury, 0 = property damage).



Conclusion

- ▶ Purpose of this project was to predict severity of car accident from sample data that I get from Coursera Capstone, This project is targeted to police officer or any interested stakeholder, Algorithm that I used was RandomBoost Algorithm which give our model accuracy of 97%.
- 



Thank You !!!