**Milestone 1**

* **Section1: Preprocessing techniques:**

1. **Dropped id as it won’t be efficient on the dataset since each ride has a different unique Id**
2. **Binary encoded cab type as there are only two types,**
3. **Tried label encoding for categorical columns but it gave a large error because the data are nominal not ordinal**
4. **Applied one hot encoding for categorical columns**
5. **In polynomial model we try to drop some features and select some according to correlation but it gave us bad results**
6. **We make some changes on data set in merge\_data(weather\_data, rides\_data)method we will explain it in the next section below.**

* **Section2: Analysis on the dataset :**

**For the weather data set we grouped the columns by location and then dropped time stamp as there are no exact similar timestamps in the taxi rides dataset, then we calculated the average weather of each location and then merged it with rides dataset twice, once as source and once as a destination, and we used the location as the merging key**

* **Section3: regression techniques:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **multivariable Linear model** | **Polynomial model** |
| **Accuracy** | 0.928917313 | 0.92891655 |
| **Train size** | 357224 rows  58 features in each row | 357224 rows  1770 features in each row |
| **Test size** | 153097 rows  58 features in each row | 153097 rows  1770 features in each row |
| **MSE of test data set in each** | 6.163785082304501 | 6.163850727158325 |
| **Training time** | It takes around 4.977044343948364 seconds | It takes around 109.91327476501465 seconds  Hint: more than one minute |

**Hint: in the two models we use the same preprocessing method as we combine the methods of preprocessing in the same file to reduce the redundancy in code and the techniques we used in preprocessing mentioned previous in this report.**

* **Screenshots:**

**A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceThere is some screenshots to explain the output**

Figure 2: polynomial model run

Figure 1: linear model run

**Chart, treemap chart

Description automatically generated**

Figure 3: the best correlation between features and we notice that it was so poor correlation

* **Conclusion :**

**Our conclusion is that the multiple linear model is more suitable in our case because it performs well and faster than polynomial mode and poly model works as well as the linear model, but it takes too much time to train so we find .**