DMT Project

November 16, 2024

[25]:
```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
```

[26]:
```
file_path="data/rideshare_kaggle.csv"
cab_rides_data=pd.read_csv(file_path)
```

[27]:
```
cab_rides_data.head()
```

[27]:

	id	timestamp	hour	day	month	\
0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	1.544953e+09	9	16	12	
1	4bd23055-6827-41c6-b23b-3c491f24e74d	1.543284e+09	2	27	11	
2	981a3613-77af-4620-a42a-0c0866077d1e	1.543367e+09	1	28	11	
3	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	1.543554e+09	4	30	11	
4	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	1.543463e+09	3	29	11	

	datetime	timezone	source	destination	\
0	2018-12-16 09:30:07	America/New_York	Haymarket Square	North Station	
1	2018-11-27 02:00:23	America/New_York	Haymarket Square	North Station	
2	2018-11-28 01:00:22	America/New_York	Haymarket Square	North Station	
3	2018-11-30 04:53:02	America/New_York	Haymarket Square	North Station	
4	2018-11-29 03:49:20	America/New_York	Haymarket Square	North Station	

	cab_type	…	precipIntensityMax	uvIndexTime	temperatureMin	\
0	Lyft	…	0.1276	1544979600	39.89	
1	Lyft	…	0.1300	1543251600	40.49	
2	Lyft	…	0.1064	1543338000	35.36	
3	Lyft	…	0.0000	1543507200	34.67	
4	Lyft	…	0.0001	1543420800	33.10	

	temperatureMinTime	temperatureMax	temperatureMaxTime	\
0	1545012000	43.68	1544968800	

```
1          1543233600           47.30          1543251600
2          1543377600           47.55          1543320000
3          1543550400           45.03          1543510800
4          1543402800           42.18          1543420800

   apparentTemperatureMin  apparentTemperatureMinTime  apparentTemperatureMax  \
0                   33.73                  1545012000                   38.07
1                   36.20                  1543291200                   43.92
2                   31.04                  1543377600                   44.12
3                   30.30                  1543550400                   38.53
4                   29.11                  1543392000                   35.75

  apparentTemperatureMaxTime
0                 1544958000
1                 1543251600
2                 1543320000
3                 1543510800
4                 1543420800

[5 rows x 57 columns]
```

[28]:
```
cab_rides_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 693071 entries, 0 to 693070
Data columns (total 57 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   id                          693071 non-null  object
 1   timestamp                   693071 non-null  float64
 2   hour                        693071 non-null  int64
 3   day                         693071 non-null  int64
 4   month                       693071 non-null  int64
 5   datetime                    693071 non-null  object
 6   timezone                    693071 non-null  object
 7   source                      693071 non-null  object
 8   destination                 693071 non-null  object
 9   cab_type                    693071 non-null  object
 10  product_id                  693071 non-null  object
 11  name                        693071 non-null  object
 12  price                       637976 non-null  float64
 13  distance                    693071 non-null  float64
 14  surge_multiplier            693071 non-null  float64
 15  latitude                    693071 non-null  float64
 16  longitude                   693071 non-null  float64
 17  temperature                 693071 non-null  float64
 18  apparentTemperature         693071 non-null  float64
 19  short_summary               693071 non-null  object
```

```
 20  long_summary                 693071 non-null  object
 21  precipIntensity              693071 non-null  float64
 22  precipProbability            693071 non-null  float64
 23  humidity                     693071 non-null  float64
 24  windSpeed                    693071 non-null  float64
 25  windGust                     693071 non-null  float64
 26  windGustTime                 693071 non-null  int64
 27  visibility                   693071 non-null  float64
 28  temperatureHigh              693071 non-null  float64
 29  temperatureHighTime          693071 non-null  int64
 30  temperatureLow               693071 non-null  float64
 31  temperatureLowTime           693071 non-null  int64
 32  apparentTemperatureHigh      693071 non-null  float64
 33  apparentTemperatureHighTime  693071 non-null  int64
 34  apparentTemperatureLow       693071 non-null  float64
 35  apparentTemperatureLowTime   693071 non-null  int64
 36  icon                         693071 non-null  object
 37  dewPoint                     693071 non-null  float64
 38  pressure                     693071 non-null  float64
 39  windBearing                  693071 non-null  int64
 40  cloudCover                   693071 non-null  float64
 41  uvIndex                      693071 non-null  int64
 42  visibility.1                 693071 non-null  float64
 43  ozone                        693071 non-null  float64
 44  sunriseTime                  693071 non-null  int64
 45  sunsetTime                   693071 non-null  int64
 46  moonPhase                    693071 non-null  float64
 47  precipIntensityMax           693071 non-null  float64
 48  uvIndexTime                  693071 non-null  int64
 49  temperatureMin               693071 non-null  float64
 50  temperatureMinTime           693071 non-null  int64
 51  temperatureMax               693071 non-null  float64
 52  temperatureMaxTime           693071 non-null  int64
 53  apparentTemperatureMin       693071 non-null  float64
 54  apparentTemperatureMinTime   693071 non-null  int64
 55  apparentTemperatureMax       693071 non-null  float64
 56  apparentTemperatureMaxTime   693071 non-null  int64
dtypes: float64(29), int64(17), object(11)
memory usage: 301.4+ MB
```

[29]:
```
missing_values = cab_rides_data.isnull().sum()
```

[30]:
```
missing_values
```

[30]:
```
id                          0
timestamp                   0
hour                        0
```

day	0
month	0
datetime	0
timezone	0
source	0
destination	0
cab_type	0
product_id	0
name	0
price	55095
distance	0
surge_multiplier	0
latitude	0
longitude	0
temperature	0
apparentTemperature	0
short_summary	0
long_summary	0
precipIntensity	0
precipProbability	0
humidity	0
windSpeed	0
windGust	0
windGustTime	0
visibility	0
temperatureHigh	0
temperatureHighTime	0
temperatureLow	0
temperatureLowTime	0
apparentTemperatureHigh	0
apparentTemperatureHighTime	0
apparentTemperatureLow	0
apparentTemperatureLowTime	0
icon	0
dewPoint	0
pressure	0
windBearing	0
cloudCover	0
uvIndex	0
visibility.1	0
ozone	0
sunriseTime	0
sunsetTime	0
moonPhase	0
precipIntensityMax	0
uvIndexTime	0
temperatureMin	0

```
temperatureMinTime                0
temperatureMax                    0
temperatureMaxTime                0
apparentTemperatureMin            0
apparentTemperatureMinTime        0
apparentTemperatureMax            0
apparentTemperatureMaxTime        0
dtype: int64
```

[31]:
```
missing_values_per_cab_type=cab_rides_data.groupby('cab_type')['price'].
↪apply(lambda x:x.isnull().sum())
```

[32]:
```
missing_values_per_cab_type
```

[32]:
```
cab_type
Lyft        0
Uber    55095
Name: price, dtype: int64
```

[33]:
```
mean_uber_price = cab_rides_data[cab_rides_data['cab_type'] == 'Uber']['price'].
↪mean()
cab_rides_data['price'] = cab_rides_data['price'].fillna(mean_uber_price)
```

[34]:
```
missing_values_per_cab_type=cab_rides_data.groupby('cab_type')['price'].
↪apply(lambda x:x.isnull().sum())
missing_values_per_cab_type
```

[34]:
```
cab_type
Lyft    0
Uber    0
Name: price, dtype: int64
```

[35]:
```
cab_rides_data['is_rain'] = cab_rides_data['short_summary'].str.
↪contains('rain', case=False).astype(int)
```

[36]:
```
cab_rides_data['datetime'] = pd.to_datetime(cab_rides_data['datetime'],␣
↪format='%Y-%m-%d %H:%M:%S')
```

[37]:
```
cab_rides_data['date'] = cab_rides_data['datetime'].dt.date
cab_rides_data['time'] = cab_rides_data['datetime'].dt.time
cab_rides_data
```

[37]:

	id	timestamp	hour	day	month	\
0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	1.544953e+09	9	16	12	
1	4bd23055-6827-41c6-b23b-3c491f24e74d	1.543284e+09	2	27	11	
2	981a3613-77af-4620-a42a-0c0866077d1e	1.543367e+09	1	28	11	
3	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	1.543554e+09	4	30	11	
4	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	1.543463e+09	3	29	11	

```
…                                       …          …   …  …    …
693066  616d3611-1820-450a-9845-a9ff304a4842  1.543708e+09    23    1     12
693067  633a3fc3-1f86-4b9e-9d48-2b7132112341  1.543708e+09    23    1     12
693068  64d451d0-639f-47a4-9b7c-6fd92fbd264f  1.543708e+09    23    1     12
693069  727e5f07-a96b-4ad1-a2c7-9abc3ad55b4e  1.543708e+09    23    1     12
693070  e7fdc087-fe86-40a5-a3c3-3b2a8badcbda  1.543708e+09    23    1     12

                 datetime          timezone            source    destination  \
0      2018-12-16 09:30:07  America/New_York  Haymarket Square  North Station
1      2018-11-27 02:00:23  America/New_York  Haymarket Square  North Station
2      2018-11-28 01:00:22  America/New_York  Haymarket Square  North Station
3      2018-11-30 04:53:02  America/New_York  Haymarket Square  North Station
4      2018-11-29 03:49:20  America/New_York  Haymarket Square  North Station
…                      …                 …                 …            …
693066 2018-12-01 23:53:05  America/New_York          West End      North End
693067 2018-12-01 23:53:05  America/New_York          West End      North End
693068 2018-12-01 23:53:05  America/New_York          West End      North End
693069 2018-12-01 23:53:05  America/New_York          West End      North End
693070 2018-12-01 23:53:05  America/New_York          West End      North End

       cab_type  … temperatureMinTime temperatureMax  temperatureMaxTime  \
0          Lyft  …          1545012000          43.68          1544968800
1          Lyft  …          1543233600          47.30          1543251600
2          Lyft  …          1543377600          47.55          1543320000
3          Lyft  …          1543550400          45.03          1543510800
4          Lyft  …          1543402800          42.18          1543420800
…           …  …                   …              …                   …
693066     Uber  …          1543658400          44.76          1543690800
693067     Uber  …          1543658400          44.76          1543690800
693068     Uber  …          1543658400          44.76          1543690800
693069     Uber  …          1543658400          44.76          1543690800
693070     Uber  …          1543658400          44.76          1543690800

        apparentTemperatureMin  apparentTemperatureMinTime  \
0                        33.73                  1545012000
1                        36.20                  1543291200
2                        31.04                  1543377600
3                        30.30                  1543550400
4                        29.11                  1543392000
…                            …                           …
693066                   27.77                  1543658400
693067                   27.77                  1543658400
693068                   27.77                  1543658400
693069                   27.77                  1543658400
693070                   27.77                  1543658400

        apparentTemperatureMax  apparentTemperatureMaxTime  is_rain  \
```

```
0                       38.07                  1544958000        0
1                       43.92                  1543251600        1
2                       44.12                  1543320000        0
3                       38.53                  1543510800        0
4                       35.75                  1543420800        0
…                         …                        …       …
693066                  44.09                  1543690800        0
693067                  44.09                  1543690800        0
693068                  44.09                  1543690800        0
693069                  44.09                  1543690800        0
693070                  44.09                  1543690800        0

                date      time
0         2018-12-16  09:30:07
1         2018-11-27  02:00:23
2         2018-11-28  01:00:22
3         2018-11-30  04:53:02
4         2018-11-29  03:49:20
…                ...       …
693066    2018-12-01  23:53:05
693067    2018-12-01  23:53:05
693068    2018-12-01  23:53:05
693069    2018-12-01  23:53:05
693070    2018-12-01  23:53:05

[693071 rows x 60 columns]
```

[38]:
```
# Create "odd_time" column
cab_rides_data['odd_time'] = cab_rides_data['time'].apply(lambda x: 1 if x.hour␣
 ↪< 6 else 0)

# Create "peak_time" column
cab_rides_data['peak_time'] = cab_rides_data['time'].apply(lambda x: 1 if (x.
 ↪hour >= 8 and x.hour <= 10) or (x.hour >= 16 and x.hour <= 19) else 0)

# Print the updated dataframe
cab_rides_data.head()
```

[38]:
```
                                     id     timestamp  hour  day  month  \
0  424553bb-7174-41ea-aeb4-fe06d4f4b9d7  1.544953e+09     9   16     12
1  4bd23055-6827-41c6-b23b-3c491f24e74d  1.543284e+09     2   27     11
2  981a3613-77af-4620-a42a-0c0866077d1e  1.543367e+09     1   28     11
3  c2d88af2-d278-4bfd-a8d0-29ca77cc5512  1.543554e+09     4   30     11
4  e0126e1f-8ca9-4f2e-82b3-50505a09db9a  1.543463e+09     3   29     11

             datetime          timezone            source    destination  \
0 2018-12-16 09:30:07  America/New_York  Haymarket Square  North Station
```

```
1 2018-11-27 02:00:23  America/New_York  Haymarket Square  North Station
2 2018-11-28 01:00:22  America/New_York  Haymarket Square  North Station
3 2018-11-30 04:53:02  America/New_York  Haymarket Square  North Station
4 2018-11-29 03:49:20  America/New_York  Haymarket Square  North Station

  cab_type  … temperatureMaxTime apparentTemperatureMin  \
0     Lyft  …         1544968800                  33.73
1     Lyft  …         1543251600                  36.20
2     Lyft  …         1543320000                  31.04
3     Lyft  …         1543510800                  30.30
4     Lyft  …         1543420800                  29.11

   apparentTemperatureMinTime  apparentTemperatureMax  \
0                  1545012000                   38.07
1                  1543291200                   43.92
2                  1543377600                   44.12
3                  1543550400                   38.53
4                  1543392000                   35.75

   apparentTemperatureMaxTime  is_rain        date      time  odd_time  \
0                  1544958000        0  2018-12-16  09:30:07         0
1                  1543251600        1  2018-11-27  02:00:23         1
2                  1543320000        0  2018-11-28  01:00:22         1
3                  1543510800        0  2018-11-30  04:53:02         1
4                  1543420800        0  2018-11-29  03:49:20         1

  peak_time
0         1
1         0
2         0
3         0
4         0

[5 rows x 62 columns]
```

[39]:
```
#sorting by datetime column
cab_rides_data = cab_rides_data.sort_values(by='datetime')
```

[40]:
```
cab_rides_data['day_of_week'] = cab_rides_data['datetime'].dt.day_name()
```

[41]:
```
# Create "is_weekend" column
cab_rides_data['is_weekend'] = cab_rides_data['day_of_week'].apply(lambda x: 1␣
 ↪if x=="Saturday" or x=="Sunday" else 0)
cab_rides_data
```

[41]:
```
                                      id     timestamp  hour  day  month  \
66422   a7b50600-c6c5-4e6c-bea9-4487344196d4  1.543204e+09     3   26     11
```

```
446073  9962f244-8fce-4ae9-a583-139d5d7522e1  1.543204e+09     3   26     11
184332  4aa68a5d-abc0-4fdf-a47f-0003617afbae  1.543204e+09     3   26     11
167114  ef8b695c-c24d-4ac1-b3fe-4aa1a7ed79f4  1.543204e+09     3   26     11
184333  89f35ef7-7129-483d-b3e6-d89afdf6946d  1.543204e+09     3   26     11
…                                        …           …   …  …    …
34918   e299c3bf-a429-4b19-af4a-ebd8e9ad74f7  1.545161e+09    19   18     12
215397  20caa061-2ded-49f8-882b-1e7eae6285ff  1.545161e+09    19   18     12
166550  7f1cbf41-2136-4e37-889d-dd0dfff02d38  1.545161e+09    19   18     12
290785  8c28dc35-c4a8-41e8-abe8-d5d65849448d  1.545161e+09    19   18     12
166551  9117e97f-c492-4964-b090-9328032e00c1  1.545161e+09    19   18     12

                  datetime          timezone                   source  \
66422  2018-11-26 03:40:46  America/New_York            North Station
446073 2018-11-26 03:40:46  America/New_York         Theatre District
184332 2018-11-26 03:40:46  America/New_York                North End
167114 2018-11-26 03:40:46  America/New_York        Boston University
184333 2018-11-26 03:40:46  America/New_York                North End
…                    …              …                        …
34918  2018-12-18 19:15:10  America/New_York       Financial District
215397 2018-12-18 19:15:10  America/New_York                  Fenway
166550 2018-12-18 19:15:10  America/New_York         Haymarket Square
290785 2018-12-18 19:15:10  America/New_York  Northeastern University
166551 2018-12-18 19:15:10  America/New_York         Haymarket Square

            destination cab_type  … apparentTemperatureMinTime  \
66422   Haymarket Square     Uber  …                 1543136400
446073         North End     Uber  …                 1543136400
184332          West End     Lyft  …                 1543136400
167114       Beacon Hill     Lyft  …                 1543136400
184333          West End     Lyft  …                 1543136400
…                    …      …  …                          …
34918   Haymarket Square     Uber  …                 1545134400
215397  Theatre District     Uber  …                 1545134400
166550          Back Bay     Uber  …                 1545134400
290785       Beacon Hill     Lyft  …                 1545134400
166551          Back Bay     Uber  …                 1545134400

       apparentTemperatureMax  apparentTemperatureMaxTime  is_rain  \
66422                   43.17                  1543186800        0
446073                  43.17                  1543186800        0
184332                  43.17                  1543186800        0
167114                  43.17                  1543186800        0
184333                  43.17                  1543186800        0
…                       …                         …      …
34918                   31.84                  1545109200        0
215397                  31.84                  1545109200        0
166550                  31.84                  1545109200        0
```

```
290785                  31.84                  1545109200        0
166551                  31.84                  1545109200        0

              date      time  odd_time  peak_time  day_of_week is_weekend
66422   2018-11-26  03:40:46         1          0       Monday          0
446073  2018-11-26  03:40:46         1          0       Monday          0
184332  2018-11-26  03:40:46         1          0       Monday          0
167114  2018-11-26  03:40:46         1          0       Monday          0
184333  2018-11-26  03:40:46         1          0       Monday          0
…              …         …       …          …            …          …
34918   2018-12-18  19:15:10         0          1      Tuesday          0
215397  2018-12-18  19:15:10         0          1      Tuesday          0
166550  2018-12-18  19:15:10         0          1      Tuesday          0
290785  2018-12-18  19:15:10         0          1      Tuesday          0
166551  2018-12-18  19:15:10         0          1      Tuesday          0

[693071 rows x 64 columns]
```

[42]:
```python
# Select features and target variable
selected_features = [
    'distance', 'surge_multiplier', 'temperature', 'humidity', 'windSpeed',
    'precipIntensity', 'is_rain', 'hour', 'day_of_week', 'is_weekend'
]
target = 'price'

# Create feature matrix (X) and target vector (y)
X = cab_rides_data[selected_features]
y = cab_rides_data[target]

# Convert categorical columns (e.g., is_rain, day_of_week) to numerical values
X = pd.get_dummies(X, columns=['day_of_week', 'is_weekend'], drop_first=True)

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)

# Check the shapes of the training and testing sets
print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
```

```
X_train shape: (554456, 15)
X_test shape: (138615, 15)
```

[43]:
```python
# Linear Regression
# Train the Linear Regression model
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)
```

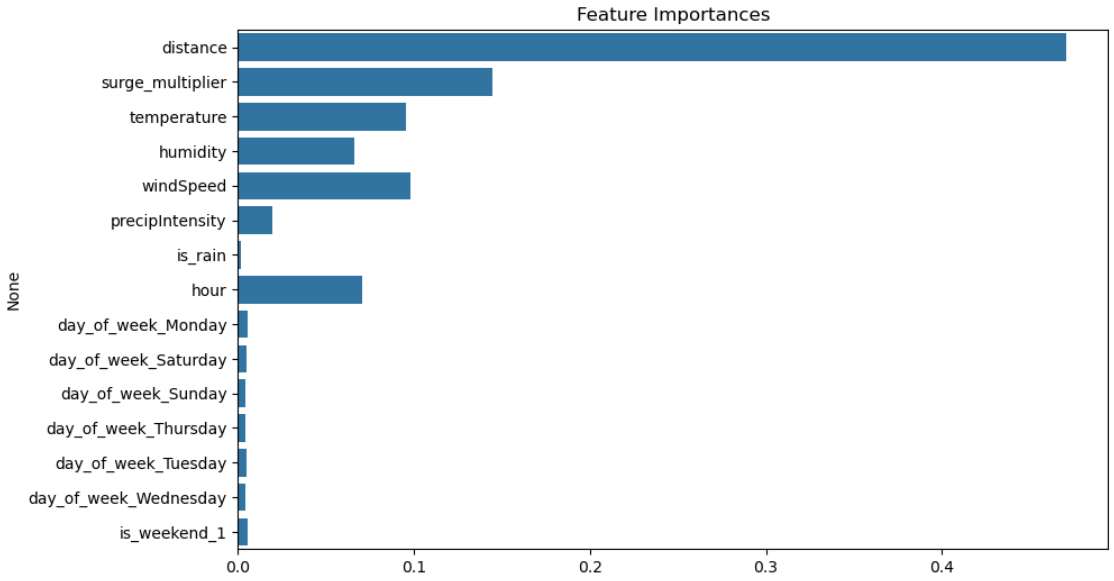
```
# Predict on the test set
y_pred_linear = linear_model.predict(X_test)

# Evaluate the model
print("Linear Regression Results:")
print("Mean Squared Error (MSE):", mean_squared_error(y_test, y_pred_linear))
print("R-squared (R2):", r2_score(y_test, y_pred_linear))
```

```
Linear Regression Results:
Mean Squared Error (MSE): 66.67380182938464
R-squared (R2): 0.16443163339397837
```

[44]:
```
# Train the Random Forest model
random_forest_model = RandomForestRegressor(random_state=42)
random_forest_model.fit(X_train, y_train)

# Predict on the test set
y_pred_rf = random_forest_model.predict(X_test)

# Evaluate the model
print("Random Forest Regressor Results:")
print("Mean Squared Error (MSE):", mean_squared_error(y_test, y_pred_rf))
print("R-squared (R2):", r2_score(y_test, y_pred_rf))
```

```
Random Forest Regressor Results:
Mean Squared Error (MSE): 80.47629683425012
R-squared (R2): -0.008543776585123952
```

[45]:
```
# Plot feature importances from Random Forest
importances = random_forest_model.feature_importances_
feature_names = X.columns

plt.figure(figsize=(10, 6))
sns.barplot(x=importances, y=feature_names)
plt.title("Feature Importances")
plt.show()
```



[46]:
```
# Compare the models
print("Model Comparison:")
print("Linear Regression - MSE:", mean_squared_error(y_test, y_pred_linear), "|␣
  ↪R2:", r2_score(y_test, y_pred_linear))
print("Random Forest Regressor - MSE:", mean_squared_error(y_test, y_pred_rf),␣
  ↪"| R2:", r2_score(y_test, y_pred_rf))
```

```
Model Comparison:
Linear Regression - MSE: 66.67380182938464 | R2: 0.16443163339397837
Random Forest Regressor - MSE: 80.47629683425012 | R2: -0.008543776585123952
```

[ ]: