Bike Sharing Demand

Pratikshya Parajuli

Ministry of Finance
Government of Nepal

Introduction

This is an automated system of renting bicycles. The process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. This project is required to combine historical usage patterns with *weather* data inorder to forecast bike rental demandin Washington DC.

Target Goal Use information available before the rental period to predict hourly bike usage for the test set.

Data Summary

- Training Set provides the data and usage of the first 19 days of each month
- Test Set provides the data from the 20th to the end of the month

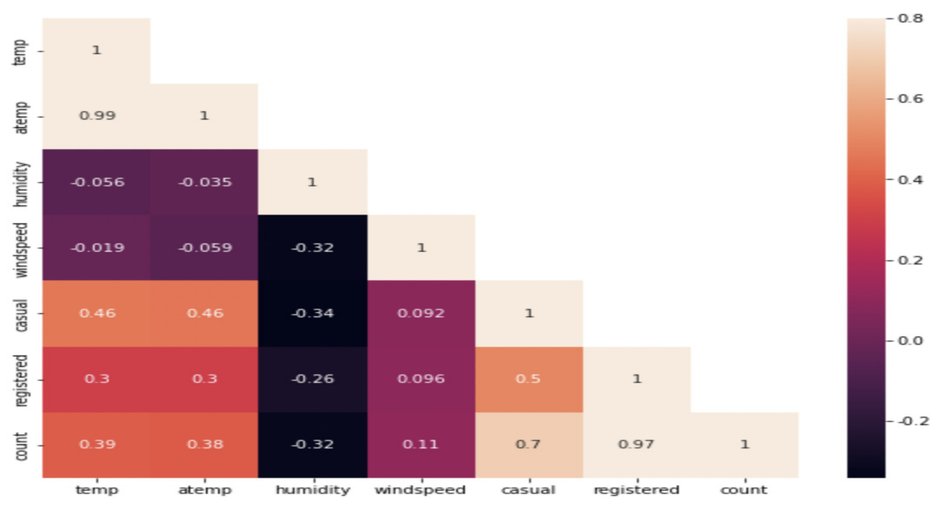
```
train_df.head()
```

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0	3	13	16
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	8	32	40
2	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0	5	27	32
3	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0	3	10	13
4	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0	0	1	1

```
train_df.dtypes
```

```
datetime      object
season         int64
holiday        int64
workingday     int64
weather        int64
temp         float64
atemp        float64
humidity       int64
windspeed    float64
casual         int64
registered     int64
count          int64
dtype: object
```

Correlation Matrix



Based on the above heatmap, we can see that some of the features have no relation with the response variable. we can drop those columns.

humidity, temp are negatively correlated with count

There is a strong correlation between temp and atemp, if both are included in the model, it will cause multicollinearity problems, so one of the features must be deleted. We remove the atemp feature because it has a weaker correlation with count than temp.

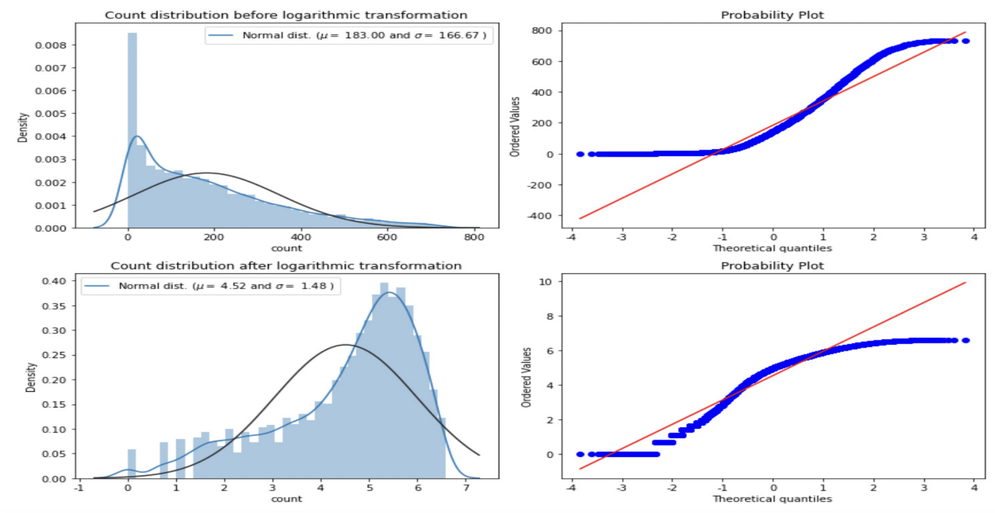
Casual, Registered are not considered and removed during model building

humidity, temp and windspeed features are considered during future modelling

fill in the zero values in the windspeed feature: usage is high when the wind speed is 0, which may be caused by null filling. Therefore, a random forest model is used here to fill in the zero values.

Target Variable Analysis

As can be seen from the figure below, the target variable count has a right-skewed distribution. Since most machine learning techniques require the dependent variable to be normally distributed, variable transformation is required. One possible solution is to log-transform the count variable after removing outlier data points. The transformed data looks much better, approximately following a normal distribution.



Model Evaluation Results

Evaluation Indicators: root mean square error is required (Root Mean Squared Logarithmic Error, RMSLE) to evaluate the quality of the model.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\log(p_i+1) - \log(a_i+1)]^2}$$

Among them, n is the number of samples in the test set, p_i is the test value, and a_i is the actual value. The smaller the root mean square error, the better the fitting effect of the data, and the closer the test value is to the actual value.

Prediction with Linear Model - Linear Regression, Ridge Regression, Lasso Regression, Logistic Regression, ElasticNet

Prediction with ensemble learning Model - Bagging Regressor, Random Forest Regressor, Gradient Boosting Regressor, AdaBoost Regressor

Model	Accuracy
Random Forest Regression	0.376319
Bagging Regression	0.395248
GBRT	0.430378
AdaBoost Regression	0.703528
Ridge Regression	1.045335
Lasso Regression	1.045453
ElasticNet Regression	1.045489
Linear Regression	1.046341
Logistic Regression	1.131105

Conclusion

Problem Definition Use information available before the rental period to predict hourly bike usage for the test set.

Prediction Algorithm Various linear model and ensemble learning algorithms are used for modelling.

Results can be further enhanced

Acknowledgement
- Tulip Lab