Executive Summary

Project Knowledge Discovery and Data Mining 2022/23

Daniel Biasiotto

May 24, 2023

Task

The dataset for the analysis was the biking dataset (link). The dataset was first preprocessed and analysed by statistical means and then a regression task predicting `cnt` was completed with different models trained on the data and tested to compare them based on results.

The metrics of `MTE` (mean squared error) and `R2` (coefficient of determination) were used to compare the models.

Tools

A `python` script was used for the analysis and training. The tools used to complete the tasks were the `matplotlib` and `seaborn` packages for data visualization and `sklearn` for the machine learning models, `numpy` for the mathematical tools and `pandas` for the data management and preprocessing.

Analysis

The data was preprocessed looking for missing values and none were found. The outliers in the weather related attributes were also characterized by a `weathersit` category of 3, the highest for weather anomaly.

Outliers in the number of customers in the biking dataset were only found in `casual` customers as expected, with no outliers in the `registered` group and in the more general `cnt`. The analysis started by considering each attribute by itself. Inspecting them through statistical means. For each attribute the 5 number summary was considered and visualized through boxplots. Then the distribution of the most interesting attributed based on the regression task was visualized to better understand the possible skew on the data. To better understand the relationships and interplay between the features a heatmap and `seaborn`'s `pairplot` were used. These plots show different distributions:

- `hum` is a slightly positively skewed gaussian
- `windspeed` is a slightly negatively skewed gaussian
- `temp` is a bimodal distribution with two peaks approximately at 0.35 and 0.65
- `cnt` is a gaussian

Additionally the plot shows a positive relationship between `temp` and `cnt`.

Attribute `instant` was removed as redundant to the task, `dteday` was converted to a simple integer `day` attribute. The `weathercond` attribute was found to be highly correlated to the target and during the optimization of the models was converted by one-hot encoding into the individual binary categories.

Using scatterplots:

- `hum` was found to be only weakly inversely correlated to `cnt`
- `temp` was found to be directly correlated to `cnt`
- `windspeed` doesn't correlate to `cnt`

Using histograms to visualize the contributions of `registered` and `casual` customers to `cnt` the mean distribution through the day was a bimodal curve with peaks around hours 8-9 and 17-18 depending on the season. `casual` customers only contributed to 20% of the total mean count. Considering the seasons the mean of the customers was highest in autumn and lowest in spring.

Using the `lmplot` of `seaborn` to try and visualize a linear relationship between the weather conditions and `cnt` showed:

- a weakly inversely to non-existent relationship with `hum`
- a weak but unclear inverse relationship with `windspeed`
- a direct relationship with `temp`
- a clearly inverse relationship with the category `weathersit`

Models

The regression models tested were:

- a simple linear model
- a ridge model
- a lasso model
- an elastic net
- a random forest

They were trained on the same data, first on the day-to-day data and then for the hourly data. For reproducibility of the test seed `111` was used by the `train-test-split` function. The `test-set` was 20% of the data.

The features used for the training were most of them except for:

- `yr`, not important for the task
- `season`, as the same information is better modeled by `mnth`
- `registered`, as part of the target of the regression
- `casual`, same of `registered`

To allow the training the `dteday` attribute was converted to a simple `day` integer attribute.

Then the models were tested again trying to improve performance. The following attribute was removed

- `atemp`, as the same information is modeled by `temp`

The categorical `weathersit` attribute was converted through the `pandas` function `get_dummies` as one-hot encoding creating 3 binary features `weathersit-1`, `weathersit-2`, `weathersit-3`.

Other attributes like `weekday` were tested through one-hot encoding but resulted in a slight lose in performance.

The results were plotted to visualize the linearity assumption and to visualize the distribution compared to the test.

Conclusions

The **Random Forest** model proved to be the most effective at predicting the target (`cnt`) by far, followed by the simple Linear Model. This was the case both in the daily and hourly training. In the daily training all measures were closer between the models, with the hourly training the Random Forest outperformed all others by a large amount. One-hot encoding the `weathersit` attribute improved the prediction slightly reducing `MTE` but mainly in the Linear Model and in the case of the day-to-day training. Interestingly the hourly training, providing many more data points to the model, improved significantly all models on the `MSE` but only the Random Forest on the coefficient of determination. See tables 1 to 4 for the results.

The same models could be trained using `casual` and `registered` attributes as targets to give further insight into the biking network.

The results of such a regression model could be used to predict the most and least congested moments in the network, for example to plan maintenance.

The data with the addition of coordinates in a city's biking network could provide interesting predictions on traffic and movement throughout the city.

Table 1: Daily results

	Linear Model	Ridge Model	Lasso Model	Elastic Net	Random Forest
MSE	1919826	1938864	1925581	2988853	1469305
R2	0.53	0.52	0.52	0.26	0.64

Table 2: Daily results optimized

	Linear Model	Ridge Model	Lasso Model	Elastic Net	Random Forest
MSE	1822903	1850090	1829576	3328053	1482230
R2	0.55	0.54	0.55	0.18	0.63

Table 3: Hourly results

	Linear Model	Ridge Model	Lasso Model	Elastic Net	Random Forest
MSE	20780	20782	20921	23581	1578
R2	0.39	0.39	0.39	0.31	0.95

Table 4: Hourly results optimized

	Linear Model	Ridge Model	Lasso Model	Elastic Net	Random Forest
MSE	20753	20753	20861	24037	1554
R2	0.39	0.39	0.39	0.30	0.95