Prediction of Bike Sharing Demand

Mei Tan Le Ping, lt528@live.mdx.ac.uk
M00724895

1.Introduction

Bicycle sharing system is a service which allows multiple users to share the use of bicycles distributed in kiosks along a city. Users can borrow a bike at a station and return it in a different station. In recent years, bike-sharing programs have become more prevalent as a result of rapid advancements in technology. Currently, there are over 500 bike-sharing programs around the world, and the number keeps on increasing. More and more people prefer to rent bikes because it is cheap, convenient, healthy and environment friendly method for short trips. Current bike share systems are automatic rental systems, which are supported by information systems.

Bike-sharing advantages is over owning a bicycle as users do not have to worry about parking issue, theft and maintenance requirements. However, the number of places where bicycles can be rented or returned are limited are the downsides. A station which will not allow users to return the bicycle, users have to search for another station that has available slots. Moreover, an empty station without bicycles to be rented will cause users to get to another station or simply use alternative mean of transportation. These inconveniences can be very frustrating for users and are sometimes the reason some of them switch to other transportation methods. Simultaneously meeting the empty bike slots and demand for bicycles is a challenging issue due to the imbalances between the return and rent rates at the stations.

Even though some bicycle balance incentive systems have been developed, some of the imbalance will still persist. In order to satisfy the user demand subject to these imbalances a fleet of light trucks transferring bicycles among stations is usually used on bicycle sharing systems. Repositioning bikes during the night when the system is almost idle is called static repositioning, meanwhile doing so during the day to cope with looming shortages is called dynamic repositioning. In order to plan a repositioning system, a demand forecast model can be used to predict which stations will be full and which ones will run out of bicycles so that the system can relocate them as needed. As a first approach to solving this problem, a data based forecast method is going to be developed to predict the total demand of bicycles for a bicycle sharing system for each hour, based on meteorological data. The study presented here is analyzing the effect of weather like average temperature, humidity, average wind speed, and weather situation, day of the week either it is holiday or working day, month, and season on the use of datasets from UCI Machine Learning repository.

2. Goals

The aim of this study is to build regression model with a small error rate to predicting fluctuations bike rental demand for both casual and registered bikers by combining historical usage patterns with the related information of weather, holiday and weekend. Machine learning which evolved from pattern recognition and artificial intelligence is that computers can learn from data through algorithms and once trained they can make data-driven predictions.

To assess the performance of algorithms, observations are usually divided into a training set and a test set. The training set is used together with the machine learning algorithm to teach the computer a model, the test set is used to evaluate how good that model can generalize by comparing predictions of the trained algorithm on the test set with the known results through a loss function.
The bicycle sharing system demand forecast problem requires of a regression algorithm, since the final value that is being forecasted is the hourly count of rented bicycles. There are multiple different machine learning regression algorithms that come from different approaches such as decision tree learning, random forest, etc. However, the similarity is that no distribution on the data is assumed and this makes the flexibility.

Machine learning methods have shown better accuracy than preexisting methods on sediment transport boosted decision trees have been found to have high prediction rates on predicting the outcome of construction litigation. For this assignment, the gradient boosted trees was chosen due to its robustness and its ability to generalize well.

3. Methodology

Based on case studies, review of appropriate predictive modelling algorithms and evaluation methods is conducted for data mining process. KNIME is used for data exploration, data mining to find the quality issues and interesting patterns of the dataset. Regression, random forest models will be built to predict the rental demand as count, and model enhancing techniques such as feature selection base or linear correlation is used to improve model’s accuracy.

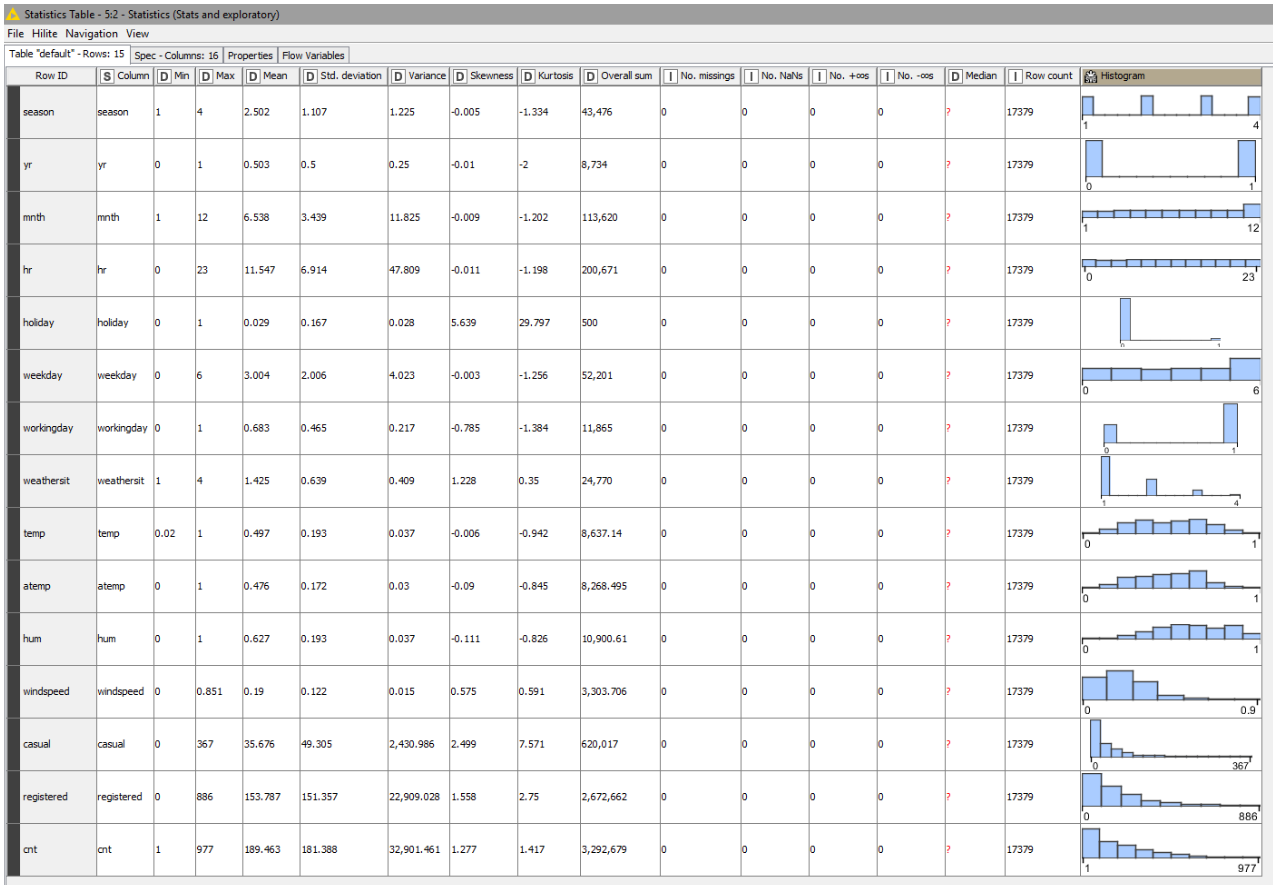
KDD (Knowledge Discovery in Databases) is the whole process that discover knowledge from data, which provides a standard process for empirical research. KDD process conations five core steps: data selection, pre-processing, transformation, data mining and Evaluation. Data mining is only an essential step of KDD. Initially, the problem and the methodologies will be used, in order to achieve the proposed goal. The first step was to acquire the data on which the model would be built by downloading data from UCI Machine Learning Repository. Then KNIME was used to visualize and explore the data. After that, an initial linear regression model was built and tested on the data making use of the previous insights. Using that model as a starting point, the model was modified and iterated on through the scientific method: hypothesizing and testing if said hypothesis made the model better at predicting the bike rentals. A final model was developed and used to draw some conclusions on the problem.

4. Data Exploration

The dataset acquired contains hourly shared bike rental information for Washington D.C. for year 2011 and 2012 and composed of 17,379 different data points with each contained 13 different input features and 3 response features. Bike rental records started from Jan 01, 2011 till Dec 31, 2012.

Variable	Description / Feature
Dteday	Date. MM/DD/YYY
Season	Type of season.1= Spring, 2 = Summer, 3 = Fall, 4 = Winter
Yr	Fiscal year. 0 = 2011, 1 = 2012
Mnth	Month. 1 = January to 12 = December
Hr	Hour. 0 = 00 hour to 23 = 23 hour
Holiday	Weather the day is holiday or not. 1 = Holiday, 0 = Not Holiday
Weekday	Weekday. 0 = Sunday to 6 =Saturday
Workingday	Whether the day is working day or not. 1 = Working day, 0 = Not working day
Weathersit	Type of weather situation. 1= Clear, 2 =Cloudy, 3 = Light Snow/Rain 4 = Heavy Snow/Rain
Temp	Temperature in Celsius, derives via(t-t_min)(t_max-t_min),T_min=-8,t_max=+39
Atemp	Feeling temperature in Celsius, derives via(t-t_min)(t_max-t_min),T_min=-16,t_max=+50(only in hourly scale)
Hum	Humidity, divided to 100(max)
Windspeed	Wind speed, divided to 67(max)
Casual	Count of casual bikers.
Registered	Count of registered bikers.
Cnt	Count of Total bikers (casual + registered)

Table1. Dataset Variables



Statistics Table - 5:2 - Statistics (Stats and exploratory)

Row ID	S Column	D Min	D Max	D Mean	D Std. deviation	D Variance	D Skewness	D Kurtosis	D Overall sum	I No. missings	I No. NaNs	I No. +∞s	I No. -∞s	D Median	I Row count	Histogram
season	season	1	4	2.502	1.107	1.225	-0.005	-1.334	43,476	0	0	0	0	?	17379	1 4
yr	yr	0	1	0.503	0.5	0.25	-0.01	-2	8,734	0	0	0	0	?	17379	0 1
mnth	mnth	1	12	6.538	3.439	11.825	-0.009	-1.202	113,620	0	0	0	0	?	17379	1 12
hr	hr	0	23	11.547	6.914	47.809	-0.011	-1.198	200,671	0	0	0	0	?	17379	0 23
holiday	holiday	0	1	0.029	0.167	0.028	5.639	29.797	500	0	0	0	0	?	17379	0 1
weekday	weekday	0	6	3.004	2.006	4.023	-0.003	-1.256	52,201	0	0	0	0	?	17379	0 6
workingday	workingday	0	1	0.683	0.465	0.217	-0.785	-1.384	11,865	0	0	0	0	?	17379	0 1
weathersit	weathersit	1	4	1.425	0.639	0.409	1.228	0.35	24,770	0	0	0	0	?	17379	1 4
temp	temp	0.02	1	0.497	0.193	0.037	-0.006	-0.942	8,637.14	0	0	0	0	?	17379	0 1
atemp	atemp	0	1	0.476	0.172	0.03	-0.09	-0.845	8,268.495	0	0	0	0	?	17379	0 1
hum	hum	0	1	0.627	0.193	0.037	-0.111	-0.826	10,900.61	0	0	0	0	?	17379	0 1
windspeed	windspeed	0	0.851	0.19	0.122	0.015	0.575	0.591	3,303.706	0	0	0	0	?	17379	0 0.9
casual	casual	0	367	35.676	49.305	2,430.986	2.499	7.571	620,017	0	0	0	0	?	17379	0 367
registered	registered	0	886	153.787	151.357	22,909.028	1.558	2.75	2,672,662	0	0	0	0	?	17379	0 886
cnt	cnt	1	977	189.463	181.388	32,901.461	1.277	1.417	3,292,679	0	0	0	0	?	17379	1 977

Figure 1. Statistics output view of all attributes.

5. Data Processing and visualisation

A preliminary data processing of checking the missing value and then attributes selection base of the linear correlation table of Figure.2. Highly correlated attributes were filtered out such as date, atemp, casual and registered biker before building any model or making any predictions. Data frame was constructed by KNIME which has the capability to detect outliers in the data based on the significance level which is shown in Figure.3 which attributes of windspeed and count with most outliers. General views of the distributions of data was shown via histogram in Figure.1

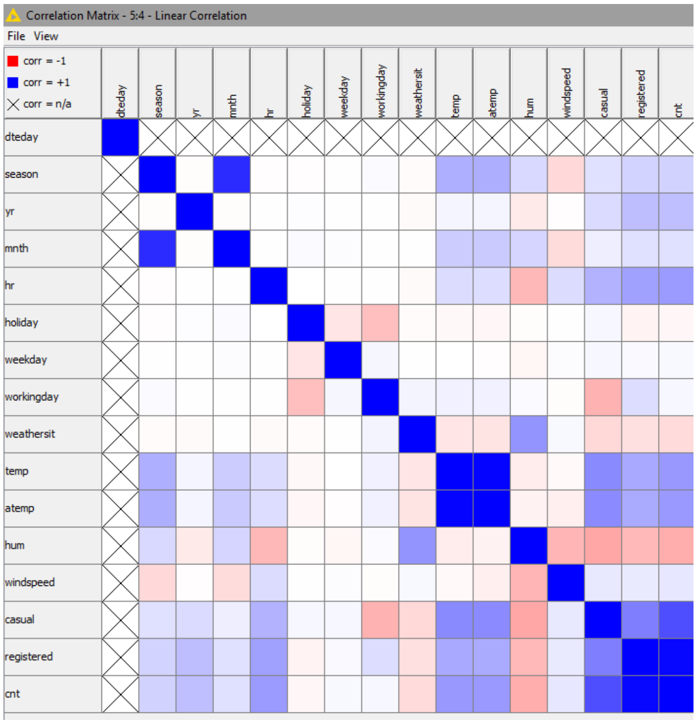
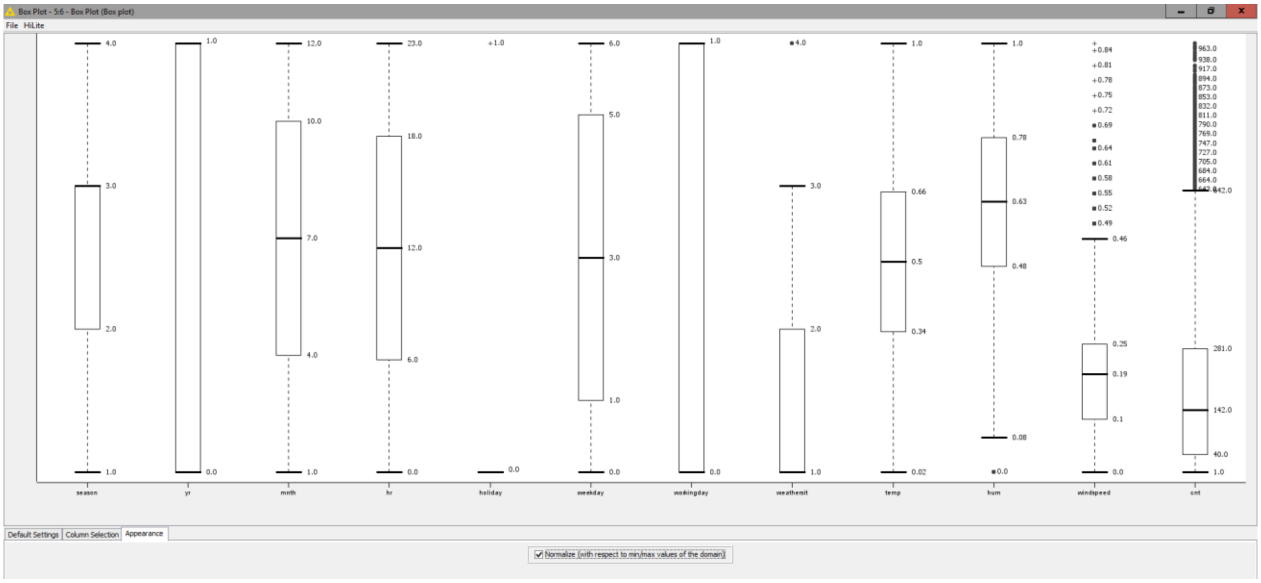


Figure 2. Linear Correlation of all the attributes

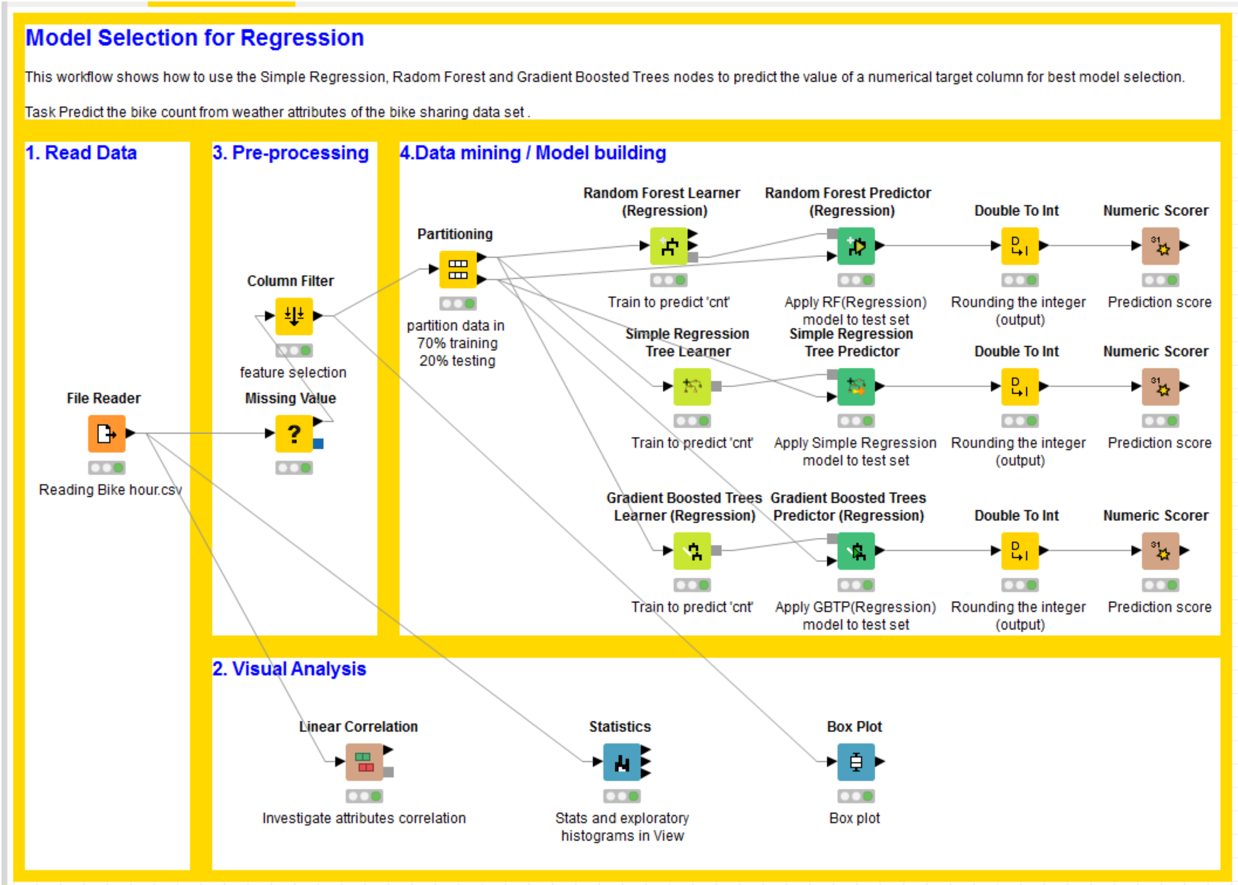


☑ Normalize (with respect to min/max values of the domain)

Figure 3. Box plot distributions of all the attributes

6. Modeling

KNIME is used for the modelling process as shown in below Figure 4. Data extracted day of year information of hourly used along with other weather metrics to predict the number of rides. Initial exploratory analysis using Statistics node revealed the significant of variables to prediction as shown in Figure1, and also shown distribution. Column filter was carried out to reduce highly correlation features. Before diving into building statistical models, dataset was partitioning into two sets, training and testing. Training set will be used to train statistical models and estimate coefficients, while testing set will be used to validate the model we build with the training set. 70% of the complete data is partitioned into training set, sampled uniformly without replacement, and 30% is partitioned in to testing set. Sampling without replacement enables the model build to extrapolate on the testing data, giving us a better sense of how statistical models perform. From the bike sharing dataset, the resulting training set contains 12165 observations and testing set contains 5214 observations.

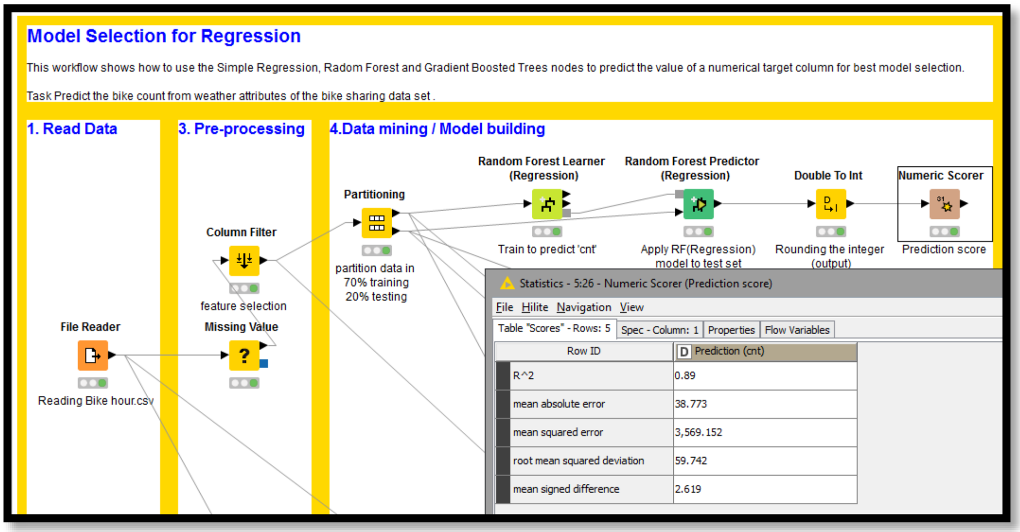


Model Selection for Regression

This workflow shows how to use the Simple Regression, Radom Forest and Gradient Boosted Trees nodes to predict the value of a numerical target column for best model selection.

Task Predict the bike count from weather attributes of the bike sharing data set .

Figure 4. Process Flow of Modeling process



Model Selection for Regression

This workflow shows how to use the Simple Regression, Radom Forest and Gradient Boosted Trees nodes to predict the value of a numerical target column for best model selection.

Task Predict the bike count from weather attributes of the bike sharing data set .

1. Read Data
3. Pre-processing
4.Data mining / Model building
File Reader
Reading Bike hour.csv
Column Filter
feature selection
Missing Value
Partitioning
partition data in 70% training 20% testing
Random Forest Learner (Regression)
Train to predict 'cnt'
Random Forest Predictor (Regression)
Apply RF(Regression) model to test set
Double To Int
Rounding the integer (output)
Numeric Scorer
Prediction score

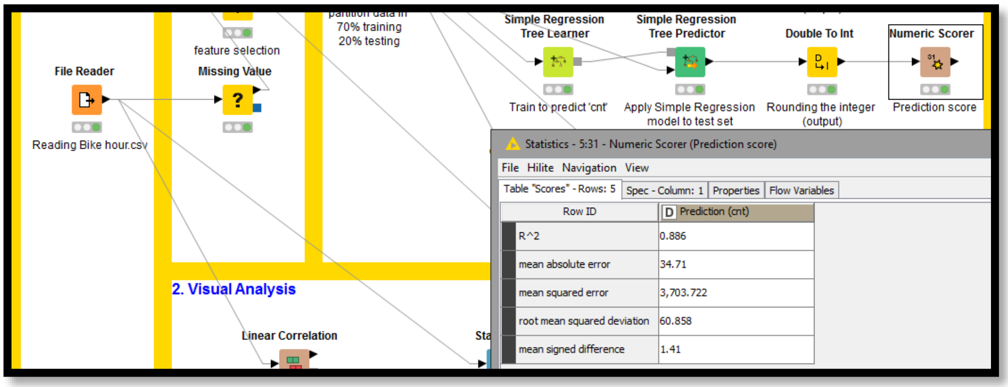
Statistics - 5:26 - Numeric Scorer (Prediction score)

File Hilite Navigation View

Table "Scores" - Rows: 5 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D Prediction (cnt)
R^2	0.89
mean absolute error	38.773
mean squared error	3,569.152
root mean squared deviation	59.742
mean signed difference	2.619

Figure 5. Prediction score of Random Forest(Regression)



partition data in 70% training 20% testing
feature selection
Missing Value
File Reader
Reading Bike hour.csv
Simple Regression Tree Learner
Train to predict 'cnt'
Simple Regression Tree Predictor
Apply Simple Regression model to test set
Double To Int
Rounding the integer (output)
Numeric Scorer
Prediction score
2. Visual Analysis
Linear Correlation
Sta

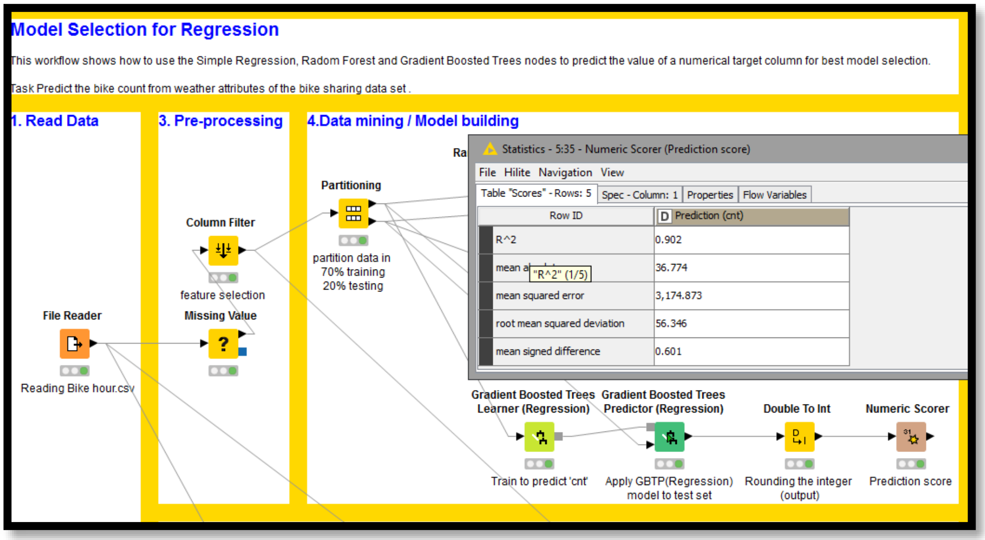
Statistics - 5:31 - Numeric Scorer (Prediction score)

File Hilite Navigation View

Table "Scores" - Rows: 5 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D Prediction (cnt)
R^2	0.886
mean absolute error	34.71
mean squared error	3,703.722
root mean squared deviation	60.858
mean signed difference	1.41

Figure 6. Prediction score of Simple Regression Tree



Model Selection for Regression

This workflow shows how to use the Simple Regression, Radom Forest and Gradient Boosted Trees nodes to predict the value of a numerical target column for best model selection.

Task Predict the bike count from weather attributes of the bike sharing data set .

1. Read Data
3. Pre-processing
4.Data mining / Model building
File Reader
Reading Bike hour.csv
Column Filter
feature selection
Missing Value
Partitioning
partition data in 70% training 20% testing
Ra
Gradient Boosted Trees Learner (Regression)
Train to predict 'cnt'
Gradient Boosted Trees Predictor (Regression)
Apply GBTP(Regression) model to test set
Double To Int
Rounding the integer (output)
Numeric Scorer
Prediction score

Statistics - 5:35 - Numeric Scorer (Prediction score)

File Hilite Navigation View

Table "Scores" - Rows: 5 | Spec - Column: 1 | Properties | Flow Variables

Row ID	D Prediction (cnt)
R^2	0.902
mean a "R^2" (1/5)	36.774
mean squared error	3,174.873
root mean squared deviation	56.346
mean signed difference	0.601

Figure 7. Prediction score of Gradient Boosted Trees

After variables was partitioning using Data Partition node. It was to train into model simple regression model, random forest regression and neural gradient boosted tree as shown in Figure.5, Figure.6 and Figure.7. These built models is connected to the predictor model for validation. And the test set 30% will be validation by the predictor node of the 3 models built as well. This will created a new column or attribute of prediction of count which will be used to validate the accuracy of models. These later used validation error to choose the best model. The output of predictor table of new attributes predictor count was in different data type as a double number whereas the target attributes count was in integer number. Thus the node of conversion double into integer were carried out for all the models before validation of models. Later, numeric score nodes were used to test the model's accuracy. The accuracy results of the prediction count of bike. Validation criterion are used for selecting the best model. As compared from the Figure 5, 6 and 7, gradient boosted tree was the best model based on its performance on validation data. The fit statistics and chosen model performance on training and validation data are reported with R-squared (R2) is a statistical measures with 90.2% accuracy.

7. Conclusion

Gradient Boosted Trees is one of the most wildly used models for data mining. Regression in machine-learning is similar to classification, with the exception of the target variable, which is a real value as opposed to a class label. In the bike-share domain, this would mean that while classification algorithms might output whether or not a station will be in overflow, balanced or shortage state, regression algorithms will output the expected number of bikes at the station. This regression assumes the target variable y on features x1, x2, x3 … xk is linear. Gradient boost tries to find the "best" line to fit the distribution of target variable with other variables, this line is the model used to predict the target variable.

Forecast methods and particularly those fueled by data and engineered using machine learning techniques have shown to be a very useful tool on the operation of a bicycle sharing system. The models described in this project take a bit of time to tune to find optimal parameters but once those parameters are optimized, they can be quickly trained and deployed. That would allow the bike share system operator to anticipate those demands using a dynamic repositioning system, easing one of the major causes of customer loss for bike sharing systems. This study is a prime example of the application of modern analytical methodologies to civil engineering and transportation problems in order to generate value by improving the experience of transport system's users.