Algorithmic Machine Learning Project

Topic : Regression
Dataset : Capital Bikeshare
Group D : Ashay Shah, Vaibhavi Mukadam, Xiaohan Zhang

Dataset is available at : http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.

Project Goal : The training set comprises the first 19 days of each month, while the test set is the 20th to the end of the month. We have to predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.

Description of dataset : The data contains 11 variables, the 'count' variable is our prediction variable. The training dataset has 10887 rows & the test dataset has 6494 rows. The discrete variables include season, holiday, workingday, weather. The discrete variables include temp, atemp, humidity, windspeed. The count variable is a summation of the registered and casual variables.

Project Objectives :

- Data preparation
- Exploratory data analysis to understand the inter-relationships between features
- Pre-Processing & Data Cleaning to detect the outliers, fix the inappropriate data points, and format the data correctly.
- Feature engineering to create new features in order to simplify & improve the model.
- Build the model : our team chose 4 models, namely Linear Regression, Random Forest, Knn & Neural Network.
- Comparing the accuracy of all 4 models and choosing the best model to predict the total count of bikes.

The 4 models are:

1 - Linear Regression : Linear Regression is one of the frequently used supervised learning techniques of machine learning. It establishes the relationship between features in a model (the independent variables) and labels (the dependent variables). Here, the optimization function or loss function is known as the residual sum of squares (RSS), which is used to define and measure the error of the model.

If we add more features to the model, its complexity increases, which again results in increasing variance and decreasing bias, i.e. overfitting. To overcome the overfitting condition, we can use regularization techniques which will decrease the model complexity by reducing the magnitudes of the coefficients. There are mainly two types of regularization techniques:
1- Ridge Regression
2- Lasso Regression
In Ridge, the cost function of the linear regression is altered by adding the penalty term (shrinkage term), which multiplies the lambda (hyperparameter) with the squared weight of each feature.
Lasso is similar to ridge regression except that the penalty term includes the absolute weight instead of the square of weights.

2- Random Forest : A Random Forest is an ensemble technique which combines multiple decision trees in determining the final output rather than relying on individual decision trees. The 2 most common techniques to perform ensemble decision trees are:
1- Bagging
2- Boosting
Bagging is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees. Average of all the predictions from different trees are used which is more robust than a single decision tree.
Boosting is where we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree. When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly. By combining the whole set at the end converts weak learners into better performing model.

3- Knn : K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure (e.g., distance functions). First, the distance between the new point and each training point is calculated. Then, The closest k data points are selected (based on the distance). The average of these data points is the final prediction for the new point.

4- Neural Network :

Using the building blocks of linear and logistic regression methods, neural networks construct more complex models for learning complicated cases.In neural networks, we first apply a matrix of weights W to the vector of input variables x to get $u = Wx$. Then we apply a nonlinear function $z = \varphi(u) = \varphi(Wx)$. The function $\varphi(x)$ is called the transfer function or the activation function. Instead of sending the outcome of the hidden layer to the output by setting $y = \varphi(Wx)$, we could instead add another layer. This means that we could take the outcome of the hidden layer, create a second layer of linear combinations and apply another set of activation functions.

Comparative analysis :

	Root Mean Square Error(RMSE)	R2 Score
Neural Network +change in parameter	87.836	0.883
Random Forest	0.430	0.907
Knn	0.5971887832979876	0.5971887832979876
Linear Regression	1.0986252878556282	-0.1985739562239146
Lasso	1.0986250591652844	-0.19857770001331088
Ridge	1.098534901791631	-0.19869383506406435

Conclusion:

- Random Forest is the best model for the bike share data for predicting count from 20th to end of the month.

Contributions :

Ashay Shah:

- Exploratory Data Analysis
- Data Cleaning & Preprocessing
- Feature Engineering
- Model Application: Random Forest, Bagging, Boosting
- Report creation & Presentation

Vaibhavi Mukadam:

- Data Preparation
- Exploratory Data Analysis
- Feature Engineering
- Model Application: Linear Regression, Ridge, Lasso, Knn
- Report creation & Presentation

Xiaohan Zhang:

- Exploratory Data Analysis
- Feature Engineering & Outlier Analysis
- Model Application: Neural Network
- Model Comparison
- Report creation & Presentation