Prepared by: Aditya Sunil Choksi (axc210030) **Date: August 7th 2023**

BUAN 6341 Applied Machine Learning
Assignment 3

A. Executive Summary

• This report offers a comprehensive analysis of the application of machine learning techniques to predict hourly bike rental demands in urban areas. The dataset used includes intricate weather metrics and hourly bike rental data from Seoul.

• A variety of methodologies were employed, including k-Means Clustering, Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Random Forest Feature Significance. These techniques were used both individually and in combination to refine and transform features for optimal model performance.

• The Stochastic Gradient Descent (SGD) was consistently applied as the learning algorithm, with the output node utilizing the Sigmoid activation function. Binary Cross Entropy was chosen as the loss function, enhancing the model's predictive capabilities.

• The report highlights the significance of feature selection and transformation. For instance, the k-Means Clustering technique was applied on features selected using Random Forest and on features transformed using PCA. The results indicated that while dimension reduction transformations like PCA compromised performance, using all features in PCA yielded comparable results to the original features.

• The bike-sharing dataset was found to be linearly distinguishable, emphasizing that altering these features doesn't significantly change outcomes. The creation of clusters, regardless of feature transformation, yielded clusters of similar sizes, which were both circular and tightly packed.

• In conclusion, the study underscores the importance of a holistic approach to feature selection, transformation, and parameter tuning in machine learning. The insights derived from this research can guide future predictions and studies in bike rental patterns, especially in relation to urban mobility and varying weather conditions.

B. Introduction

In numerous urban cities, rental bikes have been implemented to improve mobility convenience. Ensuring the timely availability and accessibility of rental bikes is crucial in reducing waiting times and addressing the challenge of maintaining a stable supply. Accurately predicting the required bike count for each hour is a vital aspect of achieving this stability. The dataset comprises weather details (such as temperature, humidity, wind speed, visibility, dew point, solar radiation, snowfall, and rainfall), along with information on the number of bikes rented per hour and date.

C. General Guidelines about Models used in the Report:

In this study, we utilized neural network models by incorporating features derived from several methodologies: 1) the k-Means Clustering technique, 2) feature refinement based on Random Forest Feature Significance, as well as transformations like Principal Component Analysis and Independent Component Analysis, and 3) a combination of k-Means Clustering attributes with features pinpointed using Random Forest Importance and Principal Components. To determine the optimal number of neurons for the first hidden layer, we implemented Cross Validation, aiming for the highest 5-fold Cross Validation results. Here are some essential aspects to consider when building the neural network:

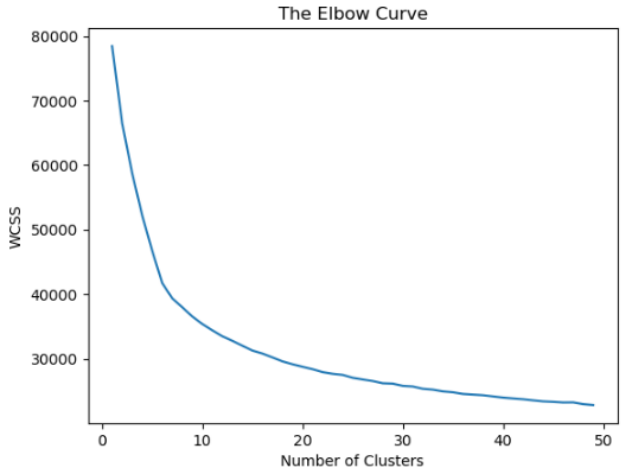
- Stochastic Gradient Descent is consistently applied as the learning algorithm (back propagation) across all tested models.
- The output node adopts the Sigmoid activation function.
- The chosen loss function is Binary Cross entropy.

D. Bike Sharing Dataset:

The Bike Sharing dataset from Seoul provides data on the hourly rental rates of bikes throughout a year, taking into account the concurrent weather conditions. The primary objective is to forecast if the hourly bike rentals surpass the median value on a given day.

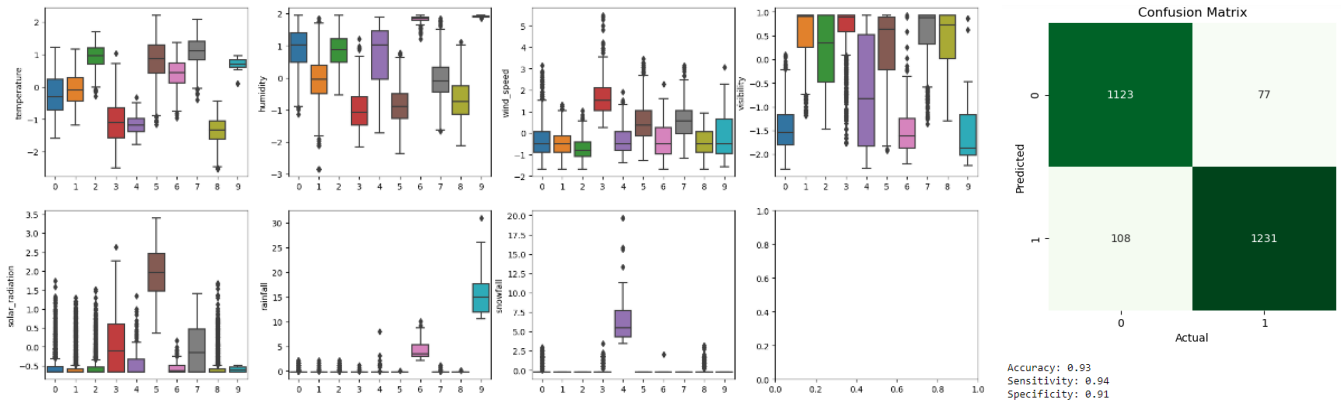
Here's a breakdown of the neural network configurations tailored for this dataset:
- Weight Modification: batch_size set at 10, total epochs set at 100
- First Hidden Layer: utilizing ReLU as the activation function, a dropout rate of 0.2, and a weight constraint defined by max_norm(3)
- Final Layer: consists of 1 neuron and employs the sigmoid activation function
- For the SGD Optimizer: a learning rate of 0.1 and momentum valued at 0.1.

E. k-Means Clustering



class	False	True
cluster		
0	1143	538
1	548	837
2	402	622
3	644	203
4	166	3
5	55	1267
6	112	4
7	59	794
8	938	111
9	18	1

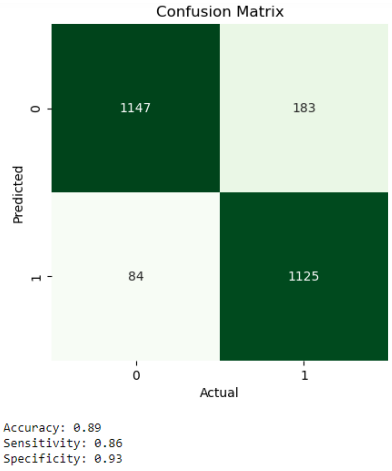
The elbow curve indicates that as we introduce more clusters, there's a decrease in the within-cluster sum of squares (WCSS). However, as we increase the cluster count, the decline in WCSS becomes more gradual. We've selected 10 as the ideal number of clusters. The subsequent boxplots reveal the data point distribution in relation to these 10 categories. Most boxplot distributions appear balanced around their median, suggesting tight clusters. Particularly, Clusters 2 and 4 exhibit elevated temperatures, increased solar exposure, average wind velocities, and diminished humidity. A significant number of data points in these clusters align with the positive class, a fact supported by the adjacent table. We've included these clusters as dummy variables (culminating in 49 features) for input into the neural network. The mean cross-validation is 0.86. The neural network learner's accuracy, sensitivity, and specificity on the test data are commendably high.



F. Dimensionality Reduction Techniques

F1. Feature Importance using Random Forest

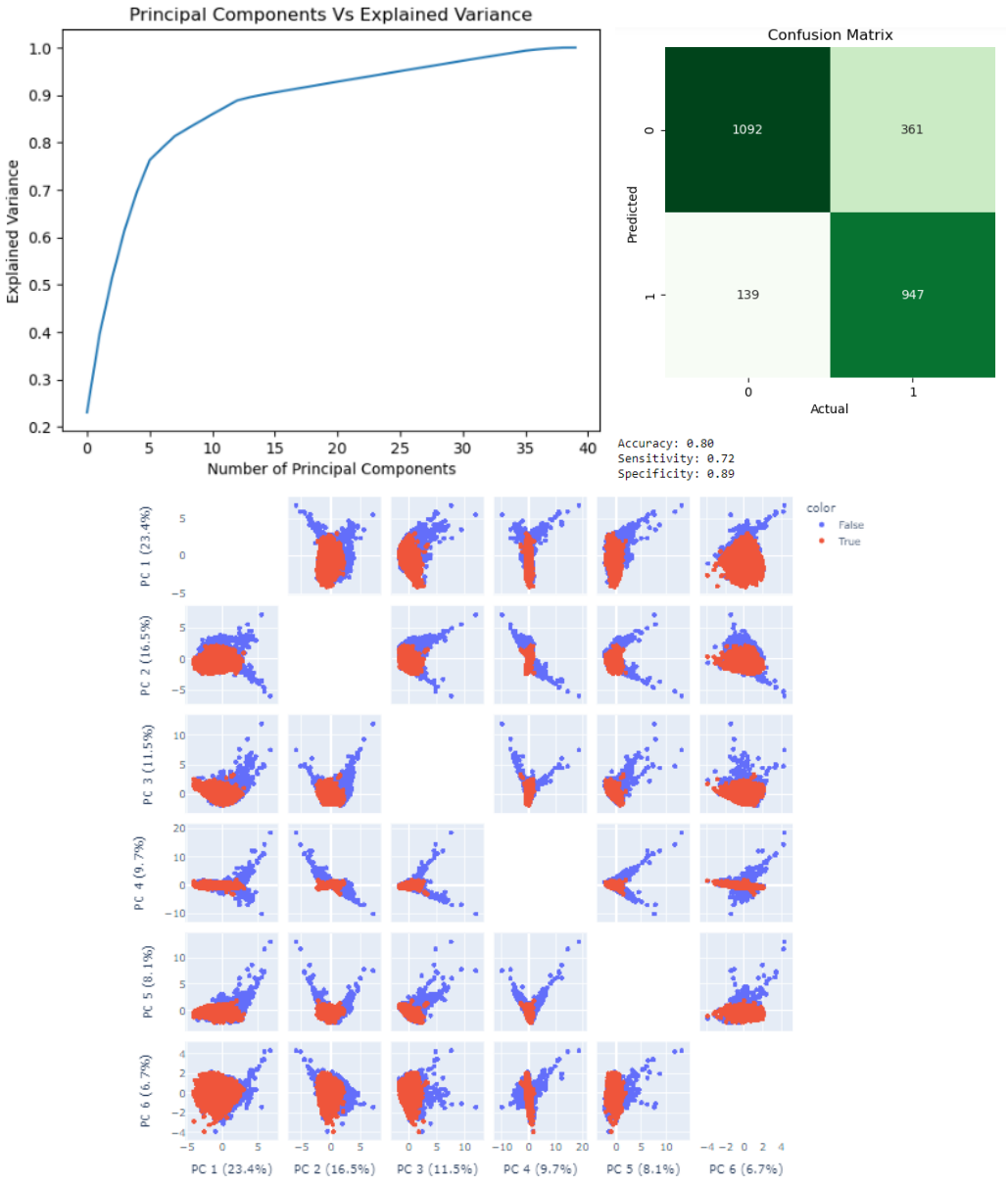
feature	importance
hour	0.220418
temperature	0.218455
season	0.187086
solar_radiation	0.096859
humidity	0.096491
visibility	0.047786
wind_speed	0.044375
day	0.042900
rainfall	0.031581
snowfall	0.010258
holiday	0.003791



Using random forests for feature selection is a method rooted in filtering-based dimensionality curtailment. Typically, variable significance is gauged by the average reduction in impurity, commonly known as Gini importance. Every tree split considers the enhancement of the split criterion, marking the significance of the variable involved in

the split. This value accumulates for each variable throughout all trees in the forest. A table presents the features ranked by their importance. For this feature selection process, only those with an importance exceeding 5% are considered. These include – hour, temperature, season, humidity, solar radiation, and visibility. In total, 29 features are provided as input to the neural network learner. Hyperparameters remain consistent with previous settings. The average cross-validation accuracy is 0.85. Since only a portion of the dataset is utilized to fine-tune the neural network, there's a notable dip in performance.

F2. Principal Component Analysis



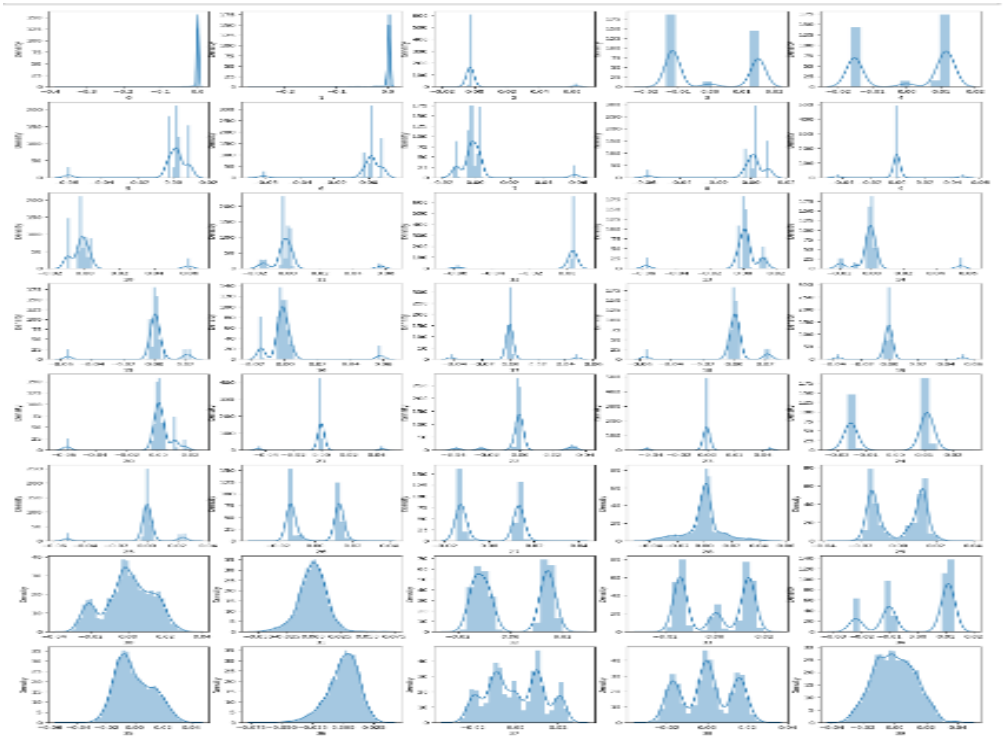
Principal Component Analysis (PCA) serves as a technique to shift features from one dimensional space to another in a linear manner. The main goal of PCA is to identify the direction that captures the most variance in the data. The number of principal components is determined by the smaller value between columns (features) or rows (entries). As we progress through the components, their capability to elucidate variance diminishes. The adjacent graph represents the accumulated variance explained by the principal components. Beyond the 13th component, the variance explanation curve tends to stabilize. The plots below demonstrate how the initial 6 principal components distinguish data points for the two classes.

The mean cross-validation accuracy stands at 0.74. Given only 13 out of 40 features were chosen, the performance isn't optimal. Nonetheless, incorporating all features pushes accuracy, sensitivity, and specificity to approach 89%.

F3. Independent Component Analysis

In contrast to PCA, Independent Component Analysis (ICA) focuses on maximizing feature independence. This involves a linear transformation from our initial feature space to a novel one, ensuring the new features are mutually independent. In ICA, the goal is to maximize mutual information between the novel features Y and the original set X. Distinctly in ICA, features aren't ranked but are ensured to be statistically independent. To evaluate the significance of features in ICA, kurtosis is employed.

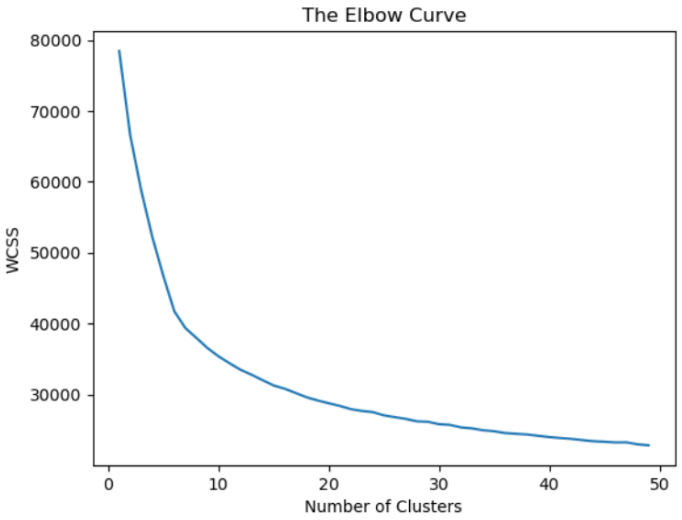
If kurtosis is positive, the variable is described as super-Gaussian or leptokurtic, which are recognized by a sharp probability density function (pdf) with pronounced tails, resembling the Laplace pdf. On the other hand, a negative kurtosis indicates that the variable is sub-Gaussian or platykurtic, characterized by a flatter pdf. The subsequent diagrams illustrate the distribution of all the independent components.

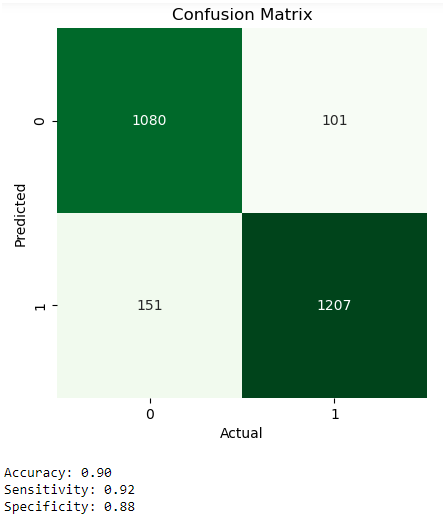


Of the 40 features, only 27 exhibit pronounced probability density functions (pdfs). Thus, these features are employed to identify local characteristics. Nevertheless, since the dataset is globally linearly separable, ICA doesn't excel in label classification.

G. Cluster Analysis on Dimensionality Reduction Algorithms

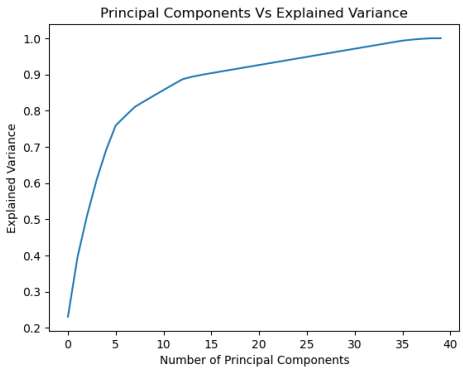
G1. k-Means on Features Selected using Random Forest

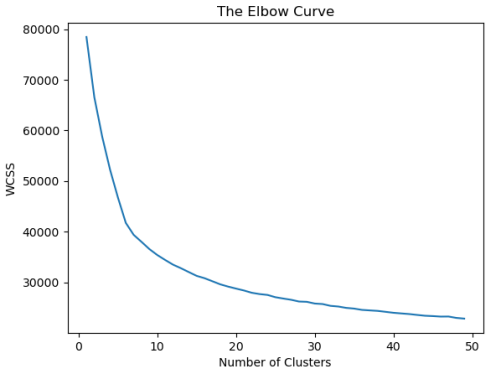




The dataset underwent k-Means clustering after feature selection based on feature significance. The appropriate cluster count was determined using the Elbow curve method, plotting the number of clusters against WCSS. A total of 10 clusters were settled upon. Features indicating cluster affiliation were transformed into dummy variables and integrated with those identified via Random Forest. This results in 38 features in total. The average cross-validation accuracy measures 0.82.

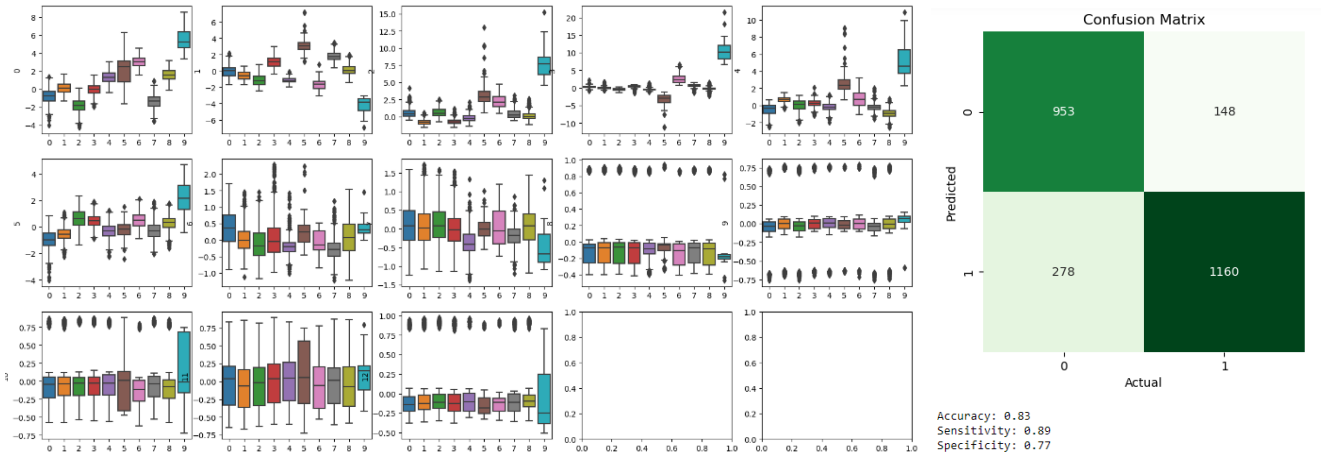
G2. k-Means on Features Transformed using PCA





The dataset is subjected to k-Means clustering after features undergo transformation through Principal Component Analysis (PCA). We chose the initial 13 PCA features because the graph stabilizes beyond the ideal value. The most fitting cluster count was discerned through the Elbow curve method, mapping the number of clusters against the within-cluster sum of squares. A conclusion was made to go with 10 clusters.

Features signifying cluster affiliation were transmuted into dummy variables and merged with the PCA-transformed features. Subsequent boxplots hint at the clusters being round in shape. The cumulative feature count stands at 24. The mean cross-validation accuracy rates at 0.73.



H. Comparison of Models

Model	5-fold CV Accuracy	Test Accuracy	Test Sensitivity	Test Specificity
k-Means Clustering	0.86	0.93	0.94	0.91
Feature Importance using Random Forest	0.85	0.89	0.86	0.93
Principal Component Analysis	0.74	0.80	0.72	0.89
Independent Component Analysis	-	-	-	-
k-Means on Features Selected using Random Forest	0.82	0.90	0.92	0.88
k-Means on Features Transformed using PCA	0.73	0.83	0.89	0.77

I. Conclusion

The bike-sharing dataset can be linearly distinguished, which means that models using the original features tend to exhibit similar behaviors. Making alterations to these features doesn't considerably alter the outcomes. Introducing transformations for dimension reduction appears to compromise performance. Yet, employing all the features in PCA would yield outcomes comparable to those using original features. The creation of clusters, whether or not feature transformation is applied, yields clusters of comparable sizes, which are both circular and tightly packed. Re-running the neural network algorithms improved the speed.