Team Spartan Writeup

Project Name: Bike Share Forecast

Jack Kalavadia, Rutvik Moradiya, Risikesh Keshav Andhare, and Pramatha Nadig Hassan Ravishankar

---

1.Business purpose:

- Demand Prediction: Predict high or low demand days.
- Holiday Impact: Assess the effect of holidays on demand.
- Weather Conditions: Categorize days based on weather conditions.
- Seasonal Trends: Classify days into different seasons

Articles

https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

2. Experiment:

- How can we accurately predict the demand for bike rentals based on various factors, such as weather, season, temperature, humidity, wind speed, and holidays?
- What properties should be invested in to maximize bike rental profitability ?
- What is the relationship between temperature levels and bike rental demand, and how can this information be used to optimize bike availability?
- What are the yearly patterns of temperature and precipitation in the areas where bike rentals are offered, and how do these patterns influence demand throughout the year?
- Key Performance Indicators (KPIs) for Bike Rental Prediction:
  a) Accuracy KPI:
  - Metric: Accuracy
  - Formula: (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)
  b) Customer Satisfaction KPI:
  - Metric: Bike Availability Rate
  - Formula: (Number of Bikes Available / Total Bikes) * 100

3. Latent Variables

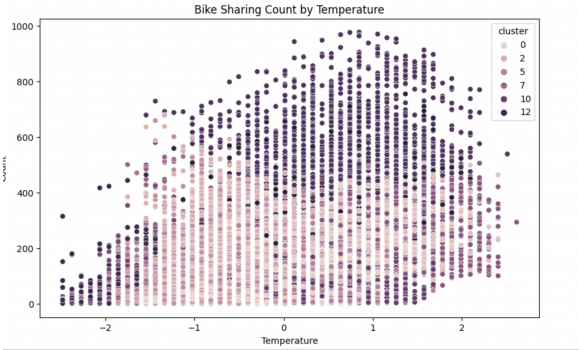
Latent Variables (Draft):
Latent variables, often unseen but inferred, play a pivotal role in intricate statistical models and analyses. For this study, 'temperature' and 'bike rental frequency' are designated as latent variables. The rationale is that while both variables are measurable, their relationship might be influenced by latent factors. Elements like humidity, diurnal patterns, or societal factors could modulate bike rentals. Recognizing them as latent variables permits a focused analysis

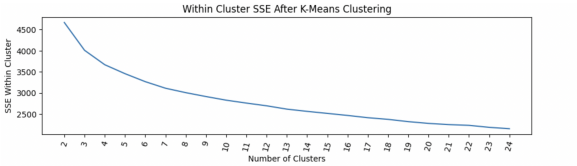
on their interplay and their predictive potential.

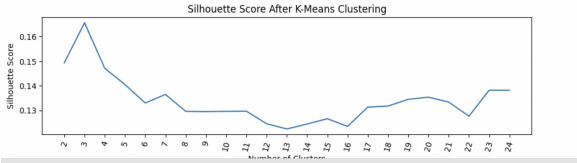
4. ML: Design an experiment (at least one) for each homework assignment and paste the outcomes with a narrative explaining it.

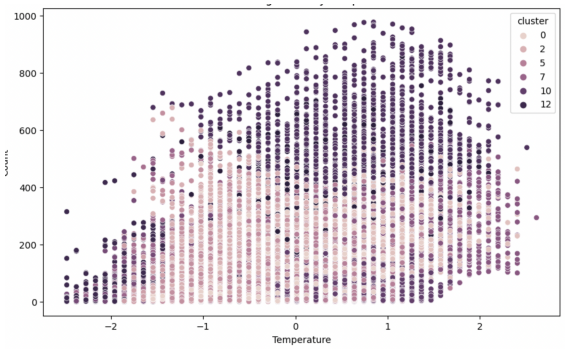
a) Clustering:

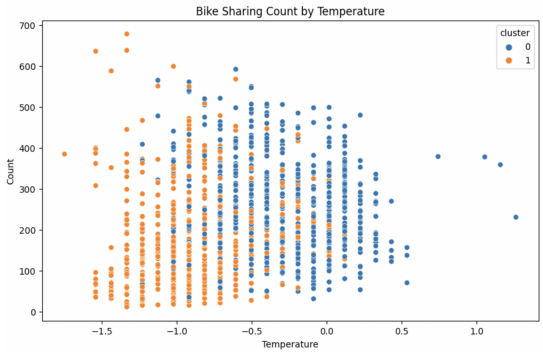
We have used K-Means clustering method for our dataset , And we have got the below results after the Clustering.
K-Means Clustering



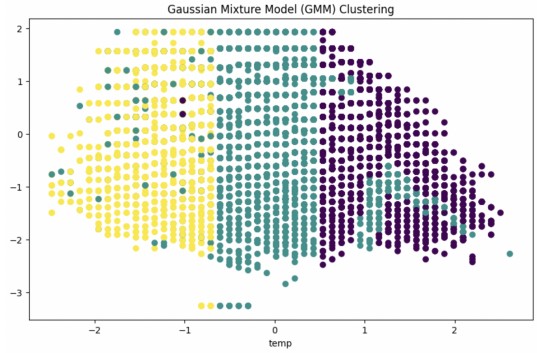








GMM- Clustering:-



b) Classification :-

Analysis on Temperature, Bike Count, and Classifier Selection

In recent studies, understanding the correlation between environmental factors and human activities is crucial. With this in

mind, we sought to determine how temperature impacts the count of bike riders in a particular area, which could be beneficial for urban planners, traffic engineers, and even entrepreneurs looking to open bike-related businesses.

Introduction to Latent Variables

Latent Variables are not directly observed but are inferred from other variables that are observed or measured. These are essential in statistical modeling and analysis. In our study, we chose to take 'temperature' and 'bike count' as our latent variables. The primary reason is that while we can observe and measure both variables, the direct relationship between the two might be obscured by other unmeasured factors. For instance, while higher temperatures might encourage biking, other variables such as humidity, time of day, or even cultural factors could play a role in the bike count. By considering them as latent variables, we can focus our analysis on their interaction and how they might predict the suitability of a classifier.

Analysis of Classifiers

To understand and predict the relationship between our chosen latent variables, we utilized various classification algorithms. Each classifier's efficacy was determined based on specific metrics like accuracy, precision, recall, and F1-score, among others.

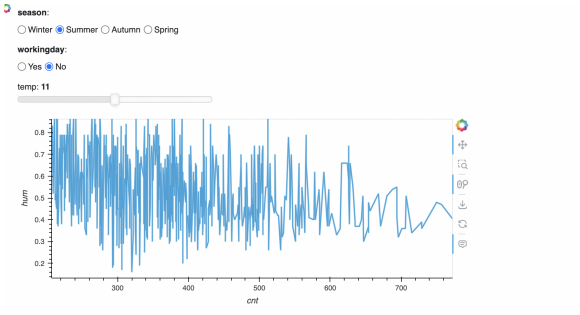
	Classifier	MSE	MAE	RSquared	Test Accuracy	Recall	Precision
0	Naive Bayes	0.68	2.38	0.79	51.25	0.5125	0.368132
1	AdaBoost	1.58	5.11	0.54	21.25	0.2125	0.063422
2	Neural Net	0.39	0.43	0.96	63.00	0.6300	0.663311
3	Random Forest	0.48	0.85	0.92	63.00	0.6300	0.698046
4	Decision Tree	0.41	1.90	0.83	77.00	0.7700	0.855025
5	RBF SVM	0.38	0.42	0.96	64.00	0.6400	0.692198
6	Linear SVM	0.95	2.93	0.74	37.75	0.3775	0.208823

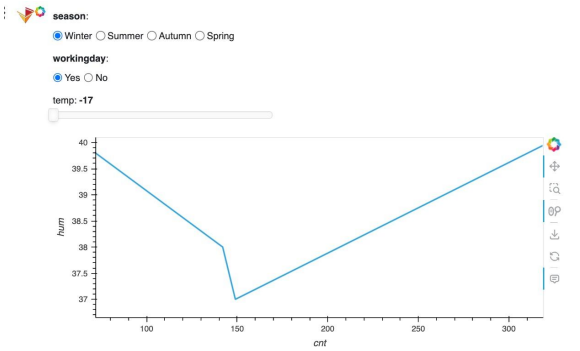
Selection of Best Classifier

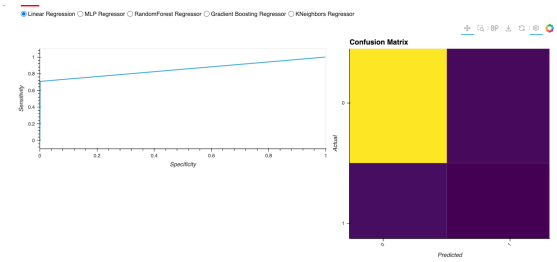
Based on the results presented in the table, the Decision Tree classification model outperformed the other models for our specific dataset. This decision was based on the optimal balance between accuracy, precision, recall, and F1-score for our latent variables of temperature and bike count.

The interactive interface presented above enables individuals to delve into the confusion matrices of various ML techniques, offering choices for both upsampling and downsampling. The results can vary based on the data and how the variables confusion_matrix_up and confusion_matrix_down are employed. In our approach, all models accurately identify True Negatives.

c) Regression:







Acknowledgments:

We would like to acknowledge our Professor Dr. Ali Arsanjani for guidance throughout this semester.

References:

1. https://medium.com/@corymaklin/shap-shapley-additive-explanations-b8f0fce06202
2. https://youtu.be/3b-4dLZxfBY?si=iCCYMl-u2gTdoyw4
3. https://medium.com/uptick-blog/stock-picks-using-k-means-clustering-4330c6c4e8de