Do we need more bikes?
Project in Statistical Machine Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this project, we aim to create a classification model that can predict whether an increase in the number of bikes is needed in Washington D.C. on a specific temporal and meteorogical conditions. We will compare various classification models which consists of Logistic Regression, Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting. The models are evaluated using Accuracy, Precision, Recall, and F1-Score metrics with 10-Fold Cross Validation. The results indicate that the Random Forest model outperforms the other models in all evaluation metrics, with an accuracy of 94.44%, recall of 94.99%, precision of 94.00%, and F1-score of 94.47%. This suggests that Random Forest is the most effective model for predicting bike demand in this context.

Number of group member: **4**

1 Problem Description

Capital Bikeshare is a 24-hour public bicycle-sharing system that serves Washington, D.C., and offers transportation for thousands of people throughout the city. The problem that arises is that there are certain occasions when, due to various circumstances, there are not as many bikes available as there are demands. In the long term, this situation will result in more people taking the car instead of the bicycle, increasing CO2 emissions in the city. To tackle this situation, the District Department of Transportation in the city wants to know if at certain hours an increase in the number of bikes available will be necessary.

In this Project, we aim to analyze whether the increase in the number of bikes is necessary or not based on the various temporal and meteorogical data provided in the dataset.

2 Data Analysis

The Training Dataset training.csv consists of 1600 randomly selected observation over the period of three years in the city of Washington D.C. The dataset contains 16 features and 1 target variable. The features are: *hour_of_day*, *day_of_week*, *month*, *holiday*, *weekday*, *summertime*, *temp*, *dew*, *humidity*, *precip*, *snow*, *snow_depth*, *windspeed*, *cloudcover*, and *visibility*. And the target variable is *increase_stock*.

2.1 Variable Types and Processing

The target variable *increase_stock* indicates whether an increase in the number of bikes is needed at a particular hour. the value *'low_bike_demand'* indicates that no increase is needed, while *'high_bike_demand'* indicates that an increase is necessary. For the analysis, we will convert these categorical values into binary numerical values, where *'low_bike_demand'* is represented as 0 and *'high_bike_demand'* as 1. Since the target variable is binary, this problem can be treated as a binary classification task.

For Binary features such as *holiday*, *weekday*, and *summertime*, they will be counted as categorical variables with values 0 and 1.

The Features *temp*, *dew*, *humidity*, *precip*, *snow*, *snow_depth*, *windspeed*, *cloudcover*, and *visibility* will be treated as numerical variables as they represent continuous measurements.

Regarding Ordinal features such as *hour_of_day*, *day_of_week*, and *month*, special attention is needed. For these features, we can't simply treat them as numerical values due to their cyclical nature. For example, after hour 23 comes hour 0 again. However, treating them as categorical variables may lead to loss of information regarding their order and cyclical patterns. Therefore, there are several possible approaches to handle them:

- One-Hot Encoding: Convert each of these features into multiple binary features, each representing a specific category. For example, *hour_of_day* would be converted into 24 binary features.
- Cyclical Transformation: Transform these features using sine and cosine functions to capture their cyclical nature. For example, for *hour_of_day*, we can create two new features:

$$\text{hour\_sin} = \sin\left(2\pi \cdot \frac{\text{hour\_of\_day}}{24}\right) [4]$$

$$\text{hour\_cos} = \cos\left(2\pi \cdot \frac{\text{hour\_of\_day}}{24}\right) [4]$$

Since One-Hot Encoding wouldn't effectively capture the cyclical nature of these features and may lead to high dimensionality, for this analysis, we will use the Cyclical Transformation approach to handle these ordinal features. [4]

After processing, the dataset will consist of 18 features and 1 target variable. Those features are shown in Table 1.

Table 1: Processed Features in the Dataset

Feature	**Type**	**Description**
hour_sin	Numerical	Sine transformation of hour of the day
hour_cos	Numerical	Cosine transformation of hour of the day
day_sin	Numerical	Sine transformation of day of the week
day_cos	Numerical	Cosine transformation of day of the week
month_sin	Numerical	Sine transformation of month of the year
month_cos	Numerical	Cosine transformation of month of the year
holiday	Binary / Categorical	Whether the day is a holiday or not (0 or 1)
weekday	Binary / Categorical	Whether the day is a weekday or not (0 or 1)
summertime	Binary / Categorical	Whether the day is in the summer time period or not (0 or 1)
temp	Numerical	Temperature in Celsius
dew	Numerical	Dew point temperature in Celsius
humidity	Numerical	Relative Humidity in percentage
precip	Numerical	Precipitation in mm
snow	Numerical	Amount of snow in the last hour in cm
snow_depth	Numerical	Accumulated snow depth in cm
windspeed	Numerical	Wind speed in km/h
cloudcover	Numerical	Percentage of cloud cover
visibility	Numerical	Distance in km at which objects or landmarks can be clearly seen and identified
increase_stock (Target)	Binary / Categorical	Whether an increase in bike stock is needed (0 or 1)

2.2 Exploratory Data Analysis

For the initial stage, we will perform Exploratory Data Analysis (EDA) to understand the distribution and trends that arises in the dataset. Including which features are more correlated with the target variable *increase_stock*.

The feature *snow* only contains zero values in all observations, therefore it will be removed from the dataset as it doesn't provide any useful information for the analysis. Upon analyzing the dataset, we found that there are no missing values in any of the features or the target variable. Therefore, no handling is required.

Using Pearson correlation coefficient, we found correlation values between each feature and the target variable as shown in Table 2.

Table 2: Ordered Correlation between Features and Target Variable

Feature	**Correlation Coefficient**
hour_of_day_cos	-0.339960
temp	0.336981
humidity	-0.308726
hour_of_day_sin	-0.308121
summertime	0.216052
month_cos	-0.169059
dew	0.132663
weekday	-0.116446
visibility	0.113443
windspeed	0.096011
month_sin	-0.092078
day_of_week_sin	-0.088152
precip	-0.059304
snowdepth	-0.047526
cloudcover	-0.045534
day_of_week_cos	-0.031473
holiday	-0.004909

As the table suggests, the feature *hour_of_day_cos* has the highest positive correlation with the target variable *increase_stock*, indicating that the time of day plays a significant role in determining whether an increase in bike stock is needed. On the other hand, the feature *holiday* has the lowest correlation with the target variable, suggesting that whether a day is a holiday or not has minimal impact on bike demand.

2.3 Imbalance in the Dataset

Upon analyzing the target variable *increase_stock*, we found that there is 1312 instances of class 0 (low bike demand) and 288 instances of class 1 (high bike demand). This indicates a significant class imbalance in the dataset [2], with class 0 being the majority class.

To address this class imbalance, we will employ the use of Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by generating synthetic samples for the minority class (class 1 in this case) by interpolating between existing minority class instances. This helps to balance the class distribution and provides the model with more representative samples of the minority class during training. [1]

The interpolation is done by selecting a minority class instance and finding its k-nearest neighbors. A synthetic sample is then created by randomly selecting one of the neighbors and interpolating between the two instances. This process is repeated until the desired balance between the classes is achieved.

3 Models and Methods

In this experiment, we will compare various classification models to determine which one performs best for predicting whether an increase in bike stock is needed using the provided dataset. The models we will consider includes Logistic Regression, Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting.

For each model, we will perform hyperparameter tuning using techniques such as Grid Search or Random Search combined with cross-validation to find the optimal set of hyperparameters that yield the best performance on the validation set.

3.1 Benchmark Model

As the benchmark model, we will use a naive model that predict each instance using stratified random sampling based on the training set's class distribution. This means that for each instance, the model will randomly assign a class label (0 or 1) based on the proportion of each class in the training data [3]. This will provide a baseline accuracy to compare the performance of more sophisticated models.

3.2 Evaluation Metrics

To evaluate the performance of each classification model, we will use several metrics including Accuracy, Precision, Recall, and F1-Score. These metrics will provide a comprehensive understanding of how well each model performs in predicting the target variable. [2]

We will also use K-Fold Cross Validation to ensure that our evaluation metrics are robust and not overly dependent on a particular train-test split.

K-Fold Cross Validation involves dividing the dataset into K subsets, using one of the subsets as the test set and the remaining K-1 subsets as the training set. This process is repeated K times, with each subset used as the test set once. The final evaluation metrics are then averaged over all K iterations to provide a more reliable estimate of model performance. For this experiment, we will use K=10. [2]

3.3 Logistic Regression

Logistic Regression is a linear model used for binary classification tasks. It models the probability of the target variable being in a particular class using the logistic function. The model estimates the coefficients for each feature, which represent the impact of each feature on the log-odds of the target variable [2].

In Logistic Regression, we use the sigmoid function to map the linear combination of features to a probability value between 0 and 1. With the formulation:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}[3]$$

Where $P(Y=1|X)$ is the probability of the target variable being in class 1 given the features X, β_0 is the intercept, and $\beta_1, \beta_2, ..., \beta_n$ are the coefficients for each feature $X_1, X_2, ..., X_n$.

The parameters of the model are estimated using Maximum Likelihood Estimation (MLE), which finds the set of coefficients that maximize the likelihood of the observed data given the model.

3.4 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a classification method that finds a linear combination of features that best separates the classes. [2].

LDA is derived from the probabilistic model which models the class-conditional distributions of the data P(Xly=k) for each class k [3]. Predictions are made by applying Bayes theorem for each training sample $x \in \mathbb{R}^d$:

$$P(y=k|x) = \frac{P(x|y=k)P(y=k)}{P(x)}[3]$$

Then we select the class with the highest posterior probability. For LDA, P(Xly=k) is modeled as a multivariate Gaussian distribution with density function as follows:

$$P(x|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right) [3]$$

where μ_k is the mean vector of class k, and Σ is the shared covariance matrix across all classes, and d is the number of features.

In LDA, we assume that the covariance matrices of all classes are equal, i.e., $\Sigma_k = \Sigma$ for all k. This reduces the log posterior to:

$$\log P(y=k|x) = -\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \log P(y=k) + Constant[3]$$

which can be simplified to:

$$\log P(y=k|x) = w_k^T x + w_{k0} + Constant[3]$$

where $w_k = \Sigma^{-1}\mu_k$ and $w_{k0} = -\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log P(y=k)$. [3]

3.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric classification method that classifies new instances based on the majority class of their K nearest neighbors in the feature space. The distance metric used to determine the nearest neighbors can be Euclidean distance, Manhattan distance, or other distance measures. [3]

KNN algorithm are usually used for unsupervised learning tasks such as clustering. However, it can also be adapted for supervised learning tasks by using the labels of the nearest neighbors to make predictions.

For Supervised learning tasks, KNN works using these following steps:

- Choose the number of neighbors K.
- For each new instance to be classified, calculate the distance between the new instance and all instances in the training dataset.
- Identify the K nearest neighbors based on the calculated distances.
- Determine the majority class among the K nearest neighbors.
- Assign the majority class as the predicted class for the new instance.

For the distance metric, we will use Manhattan distance, which is defined as:

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i|[3]$$

where p and q are two instances in the feature space, and n is the number of features.

3.6 Random Forest

Random Forest is an ensemble learning method that consists of multiple decision trees. Each tree is trained on a random subset of the training data and a random subset of features. The final prediction is made by aggregating the predictions from all trees. The method to aggregate the results in this experiment will use majority voting. [3]

Decision trees in principle work by recursively splitting the data based on feature values to create branches that lead to leaf nodes representing class labels. The splits are chosen based on the feature that maximizes the information gain or minimizes the impurity at each node. For the metric to measure impurity, we will use Entropy, which is defined as:

$$H(X) = -\sum_{i=1}^{c} p_i \log_2(p_i)[3]$$

where p_i is the proportion of instances belonging to class i in the node, and c is the number of classes.

Decision trees can be prone to overfitting, especially when they are deep and complex. To mitigate this, we can use Random Forest.

Random Forest Algorithm, works using these following steps:

- For each tree in the forest:
  - Randomly sample the training data with replacement (bootstrap sampling).
  - Randomly select a subset of features to consider for splitting at each node.
  - Train a decision tree on the sampled data using the selected features.
- For making predictions:
  - For each new instance, pass it through each tree in the forest to obtain the predicted class.
  - Aggregate the predictions from all trees using majority voting to determine the final predicted class.

Due to the randomness introduced in the training process, Random Forests are less prone to overfitting compared to individual decision trees and often achieve better generalization performance. [3]

3.7 Gradient Boosting

Gradient Boosting is an ensemble learning method that builds a series of weak learners in a sequential manner. Each weak learner is trained to correct the errors made by the previous learners. The final prediction is made by combining the predictions from all weak learners. [3]

Gradient Boosting Algorithm works using these following steps:

- Initialize the model with a constant value, typically the mean of the target variable.
- For each iteration $m = 1$ to M:
  - Compute the pseudo-residuals, which are the negative gradients of the loss function with respect to the current model's predictions.
  - Train a weak learner (e.g., decision tree, logistic regression) on the pseudo-residuals.
  - Compute the optimal step size (learning rate) for the weak learner.
  - Update the model by adding the weighted predictions of the weak learner to the current model.
- For making predictions:
  - For each new instance, pass it through all weak learners and sum their weighted predictions to obtain the final predicted value.

Since the problem we're working on is a binary classification task, we will use logistic loss as the loss function for Gradient Boosting. The logistic loss is defined as:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \text{ [3]}$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of instances. For the weak learners, we will use decision trees with a maximum depth of 5.

3.8 Hyperparameter Tuning

For each classification model, we will perform hyperparameter tuning using Grid Search combined with 10-Fold Cross Validation to find the optimal set of hyperparameters that yield the best performance on the validation set. The hyperparameters to be tuned for each model are as follows:

Grid Search works by exhaustively searching through a specified subset of hyperparameters for each model. For each combination of hyperparameters, the model is trained and evaluated using cross-validation. The combination that yields the best average performance across the folds is selected as the optimal set of hyperparameters. [2]

4 Experiment and Results

After performing the experiments using the described models and methods. We obtained the results as shown in Table 3.

Table 3: Results of Classification Models (rounded to 4 decimal places)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9037	0.9065	0.9037	0.9043
Gradient Boosting	0.8994	0.9015	0.8994	0.8999
Logistic Regression	0.8100	0.8696	0.8100	0.8271
LDA	0.7956	0.8676	0.7956	0.8155
K-Nearest Neighbors	0.7675	0.8233	0.7675	0.7863
Benchmark Model	0.4781	0.6889	0.4781	0.5375

Using Grid Search with 10-Fold Cross Validation, we found the optimal hyperparameters for each model is as follows:

- **Logistic Regression**: Regularization strength $C = 1.0$, Solver type using Liblinear, and Penalty type using Ridge Regularization (L2).
- **LDA**: Solver Type using SVD (Singular Value Decomposition)
- **K-Nearest Neighbors**: Number of neighbors $K = 3$, Distance metric using Manhattan distance, and Weighting type using the inverse of their distance.
- **Random Forest**: Number of trees $n\_estimators = 250$, Maximum depth of each tree $max\_depth = 15$, Minimum samples per leaf $min\_samples\_leaf = 1$, Minimum samples per split $min\_samples\_split = 2$, and Criterion using Entropy.
- **Gradient Boosting**: Number of estimators $n\_estimators = 100$, Learning rate = 0.1, Maximum Depth for each estimator = 10, and Subsample = 0.6

Based on the results, we can see that the more complex models such as Random Forest and Gradient Boosting outperformed the simpler models like Logistic Regression, LDA, and KNN.

Random Forest achieved an accuracy of 90.37%, indicating that it correctly classified a high percentage of instances in the dataset. The F1-Score of 90.43% also suggests that the model has a good balance between precision and recall, suggesting that it is effective in identifying both positive and negative instances. Gradient Boosting also performed comparably well, achieving an accuracy and F1-Score of 89.94% and 89.99% respectively, which is slightly lower than Random Forest but still significantly better than the other models.

If we compare the simpler models such as Logistic Regression, LDA, and KNN, we can see that they achieved lower accuracy and F1-scores. Logistic Regression achieved an accuracy of 81.00%, LDA achieved 79.56%, and KNN achieved 76.75%. This indicates that these models may not be as effective in capturing the complex relationships in the dataset compared to the ensemble methods.

Due to it's superior performance, for production deployment, we recommend using either the Random Forest or Gradient Boosting model for predicting whether an increase in bike stock is needed.

5 Conclusion

Based on the analysis and experiments conducted in this project, we can conclude that the Random Forest is the most effective model for predicting whether an increase in bike stock is needed in Washington D.C. The model achieved the highest accuracy, recall, precision, and F1-score among all the models evaluated.

References

[1] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18 (17):1–5, 2017. URL `http://jmlr.org/papers/v18/16-365.html`.

[2] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022. URL `https://smlbook.org`.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] D. Radečić. How to handle cyclical data in machine learning, 2020. URL `https://towardsdatascience.com/how-to-handle-cyclical-data-in-machine-learning-3e0336f7f97c`.