CARS_CLASS CLASSIFICATION

ABSTRACT

In terms of the travel demand prediction from the household car ownership model, if the imbalanced data were used to support the transportation policy via a machine learning model, it would negatively affect the algorithm training process. In other words, the number of members of the minority class is lower than the rest of the answer classes. The result is a bias in data classification. Consequently, this research suggested balancing the datasets with cost-sensitive learning methods, including decision trees, k-nearest neighbors (kNN), and naive Bayes algorithms. Before creating the 3-class model, a k-folds cross-validation method was applied to classify the datasets to define the true positive rate (TPR) for the model's performance validation. The outcome indicated that the kNN algorithm demonstrated the best performance for the minority class data prediction compared to other algorithms. It provides TPR for rural and suburban area types, which are region types with very different imbalance ratios, before balancing the data of 46.9% and 46.4%. After balancing the data (MCN1), TPR values were 84.4% and 81.4%, respectively.

KEYWORDS

Logistic Regression, Support Vector Machine, Decision trees, k-nearest neighbors (kNN), cross-validation.

AIM

The main aim of this project is to predict the class of a used car based on various features.

The solution is divided into the following sections:

- Data understanding and exploration.
- Data cleaning.
- Data preparation.
- Building model
- Conclusion.

INTRODUCTION

Data classification is an analysis method used to define data patterns, classification models, and classification rules. This method predicts different data types, either present or future, such as travel demand predictions. Several minor models were used, including the household car ownership models, trip generation models, tour generation models, trip distribution models, travel time choice models, and travel route choice models, with either trip or tour used as the unit of analysis. There are several techniques for data classification, e.g. the decision tree (DT) presenting different logical conditions; k-nearest neighbors (kNN) used for the mathematic calculation to and distance or weight; and naive Bayes used to the probability in the training data. The selection for a high performing technique should rely on the parameters indicating the data classification performance, e.g. accuracy, precision, recall, F1-score. Still, these techniques do not work well on every dataset. For example, some work more effectively on the balanced data than on the imbalanced one; the at data contains the class-es with a similar number of datasets. The imbalanced data has courses

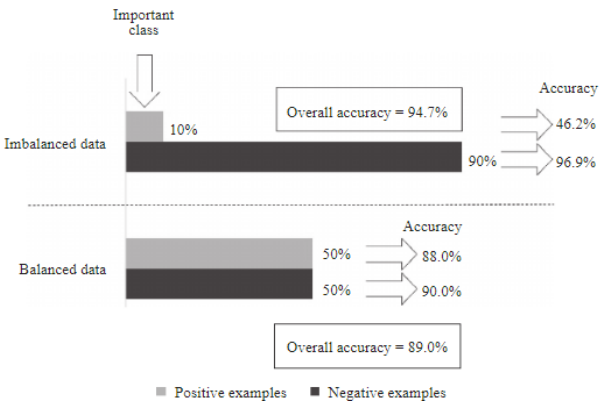
with a different number of datasets. At this point, the imbalanced data classification becomes a thought-provoking issue because some of the minority classes include either significant or outstanding data. Consequently, for more effective data analysis, the model's performance to classify the minority class needs to be improved be-fore algorithm training with suitable parameters for the imbalanced data. In the imbalanced data, the numbers of each class would be completely different. This imbalanced class is a critical issue often found in the research fields of medical science, marketing, banking, and production industry . However, it is still rare in transportation planning, especially in using the data with the machine learning model, which are popular and state-of-the-art approaches, to predict the household car ownership. Due to the problem, several methods have been purposively invented to these imbalanced data at a data level and an algorithm level to improve the minority class. Precisely at a data level, the imbalanced data could be solved via sampling techniques. Meanwhile, at an algorithm level, the algorithm's performance would be improved with any helpful technique during the data training process to effectively predict the unseen data while testing the model, such as cost-sensitive learning methods (CSL). The classification performance at both levels was similar. In increasing data, CSL methods performed better than the sampling methods. Consequently, this research aimed to improve the minority class with a cost matrix table with two categories. This research proposed a useful technique to improve the algorithm's performance to classify the household car ownership demand model with the 3-class problem. The study used CSL methods to solve the imbalanced data with its negative effect on the classification performance of the minority class, and the feature section, a feature-level data management technique, to and the ten parameters with the optimal weight. Finally, the data classification performance would be affirmed by the true positive rate (TPR), F1-score, accuracy, false negative rate (FNR), and false positive rate (FPR).

CLASS DISTRIBUTION BALANCING

This section will explain the problem that might exist due to the imbalanced data distribution in each target class and the classification performance indicators for the imbalanced data. The final part is a review of the CSL methods.

The class imbalance problem

The imbalanced data can be practically seen as unequal numbers of samples in each target class, with most classification problems in research with two categories, as seen in Figure



Methodology

The project deals with used cars. The project's methodology is as follows: Proposed Methodology

After data collection, the dataset was pre-processed to remove samples that have missing values, remove the non-numerical part from numerical attributes, convert categorical values into numerical (if needed), fix any discrepancies in the units, as well as removing attributes that don't affect the price evaluations if needed to reduce the complexity of the model. Data Understanding and preparation are essential part of building a model as it gives insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have a deeper understanding of the quality of the data, in terms of outliers and the skewness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

- NULL cells conventions
- Missing values
- Encoding
- Normalization Pre-processing
- Logistic Regression
- Random Forest Regressor
- Accuracy
- MSE
- MAE
- RMSE

Afterward when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict the prices and values of used cars. In this study, three models are proposed to be built using the Logistic Regression model technique, Random Forest Regressor, and Bagging Regressor. Firstly, the data was portioned into sections for training and another part for testing, portioning percentages can be tested with different ratios to analyze different results. All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). From all, the Random Forest Regressor outperformed.

Machine learning

The goal of machine learning (ML) is to help a computer learn without being explicitly instructed to do so by means of mathematical models of data. Artificial intelligence (AI) is a subset of machine learning. Data is analyzed using algorithms to identify patterns, which are then used to create predictive models. Like humans, machine learning becomes more accurate with more data and experience. With machine learning, you can adapt to situations where

data is constantly changing, the nature of the request or task is shifting, or coding a solution isn't feasible.

Machine Learning Categories Supervised and unsupervised learning are commonly used types while reinforcement is a sequential decision-maker technique. The main categories of supervised and unsupervised machine learning are:

- Machine learning Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Supervised Learning Working and details of some famous supervised machine learning algorithms which are used in this project are Logistic Regression: Whenever the dependent variable is non-numerical (categorical) and the class should be predicted, not classified, the logistic regression algorithm needs to be abandoned. Machine learning technique Logistic Regression is commonly used to classify binary data. The function of Logistic Regression is to optimize results based on various datasets. To predict results, the default label class is always employed, but the results and probability are always calculated after all categorical values have been converted into numerical values and all data has been normalized. Logistic regression, also referred to as sigmoid regression, was designed by statisticians to explain the properties of the population increasing in the ecological study, growing fast, and maxing out on the capability to wear out the surroundings. With an S-shaped curve, any real-valued range can be mapped right into a number between zero and 1, but not precisely at the limit of 1. 1 1 + e^{-x} Equation 1 Sigmoid equation Where e is the base of the logarithms (Euler's wide variety or the EXP() characteristic on your spreadsheet) and price is the real numerical price which needs to be transformed. While the equation of regression in which intercept and slope are integrated is as follows: $y = mx + c$ Equation 2 Regression equation Below is a generalized equation for Multivariate regression model: y = β0 + β1.x1 + β2.x2 +….. + βn.xn There are few steps involved in generating the regression beginning with feature selection, normalizing features, select loss function and hypothesis, set hypothesis parameters, and minimizing the loss function, and finally testing the function of the data. Random Forest Regressor: Random Forest is already revealing that it creates a forest and then somehow randomizes it. It builds the forest through the ensemble of Decision Trees and most of the time trains it using a method called the Bagging Method. Since it uses the ensemble method, the result is improved. Decision tree and bagging classifier hypermeters are the same. Each feature in the tree can be made random simply by adding thresholds.

Description of attributes

- Comp: Compactness
- Circ: Circularity
- D.Circ: Distance Circularity
- Rad.Ra: Radius ratio
- Pr.Axis.Ra: pr.axis aspect ratio
- Max.L.Ra: max.length aspect ratio
- Scat.Ra: scatter ratio
- Elong: elongatedness
- Pr.Axis.Rect: pr.axis rectangularity
- Max.L.Rect: max.length rectangularity
- Sc.Var.Maxis: scaled variance along major axis

- Sc.Var.maxis: scaled variance along minor axis
- Ra.Gyr: scaled radius of gyration
- Skew.Maxis: skewness about major axis
- Skew.maxis: skewness about minor axis
- Kurt.maxis: kurtosis about minor axis
- Kurt.Maxis: kurtosis about major axis
- Holl.Ra: hollows ratio

Importing Libraries

```
[1] #Importing Libraries
    import pandas as pd
    import numpy as np
    import seaborn as sns
    import matplotlib.pyplot as plt
```

Loading Datasets

```
[2] #Importing Dataset
    data=pd.read_csv('cars_class.csv')
```

Dimension of the set

```
#shape of the dataset
data.shape
```

```
(719, 20)
```

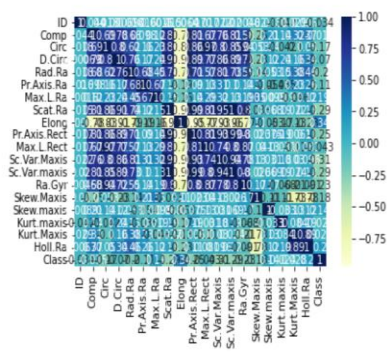
```
[4] #display the dataset
    data.head()
```

	ID	Comp	Circ	D.Circ	Rad.Ra	Pr.Axis.Ra	Max.L.Ra	Scat.Ra	Elong	Pr.Axis.Rect	Max.L.Rect	Sc.Var.Maxis	Sc.Var.maxis	Ra.Gyr	Skew.Maxis	Skew.r
0	1	88	39	70	166	66	7	148	44	19	134	167	332	143	69	
1	2	85	35	64	129	57	6	116	57	17	125	138	200	123	65	
2	3	91	41	84	141	57	9	149	45	19	143	170	330	158	72	
3	4	102	54	98	177	56	10	219	31	25	171	219	706	223	72	
4	5	87	39	74	152	58	6	151	44	19	136	174	337	140	70	

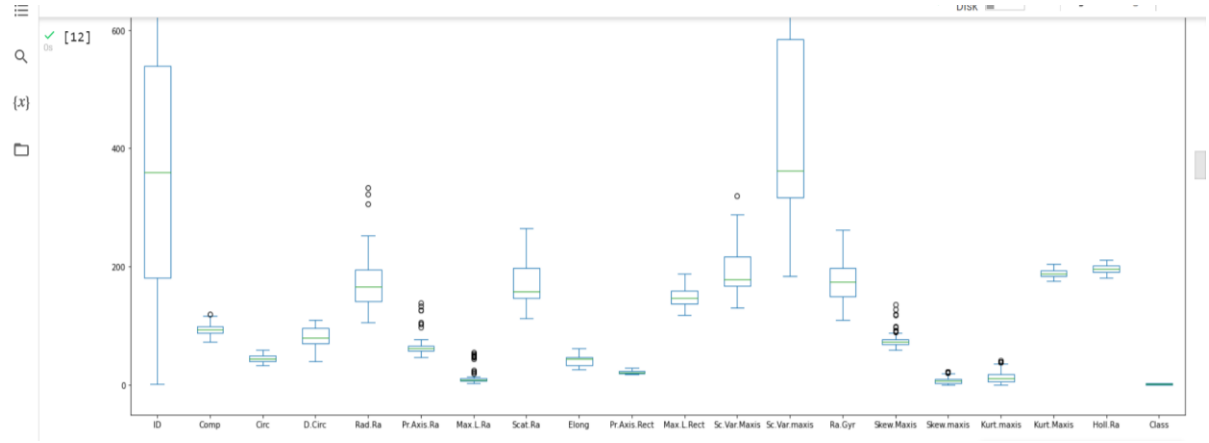
Data Correlation

```
[10] correlation=data.corr()
```

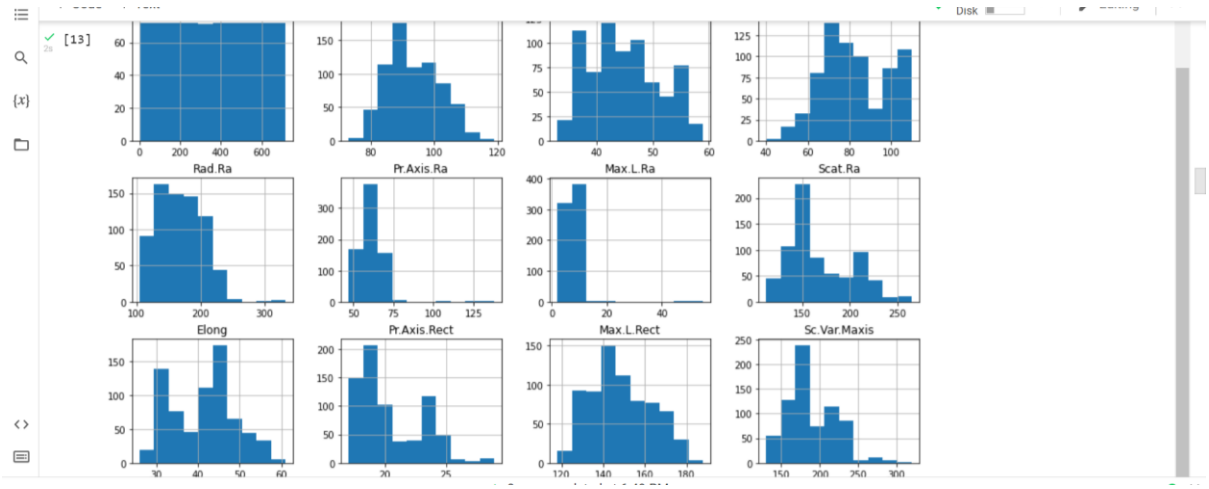
```
sns.heatmap(data.corr(), cmap="YlGnBu", annot=True);
```



Box plot to understand spread and outliers



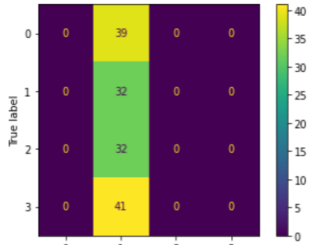
Histogram



Confusion Matrix for Logistics Regression

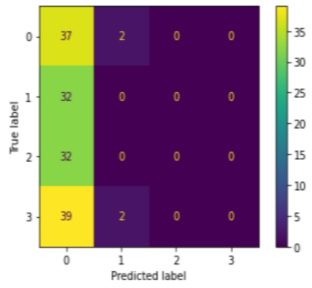
Confusion Matrix for SVC

```
[29] from sklearn.metrics import plot_confusion_matrix
     plot_confusion_matrix(svc,X_test,y_test)

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated; Function `plot_cor
  warnings.warn(msg, category=FutureWarning)
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:444: UserWarning: X has feature names, but SVC was fitted without feature names
  f"X has feature names, but {self.__class__.__name__} was fitted without"
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7fcd1d44d550>
```



Confusion Matrix for KNN

```
[35] from sklearn.metrics import plot_confusion_matrix
     plot_confusion_matrix(knn,X_test,y_test)

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated; Function `plot_cor
  warnings.warn(msg, category=FutureWarning)
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:444: UserWarning: X has feature names, but KNeighborsClassifier was fitted without feature names
  f"X has feature names, but {self.__class__.__name__} was fitted without"
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7fcd1d1e6f50>
```



Conclusion

Using data mining and machine learning approaches, this project proposed a used car price prediction. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, RandomizedSearch CV came out on top with 95% accuracy.