CASE STUDY REPORT ON SHOP CUSTOMER DATSET

AKSHAYA M S,
M.Sc. Computer Science (Data Analytics)
Rajagiri College of Social Sciences

Abstract

The Shop Customer Dataset is a collection of data which contains information about the customers who visit a particular shop. The dataset contains various features such as income, profession, income, spending score etc. which shows how much a customer spends in the shop. This dataset contains around 2000 records. This dataset can be used by businesses to analyse and gain insights about customer behaviour and their preferences in order to develop effective marketing strategies. This can help businesses to increase sales and also can improve customer satisfaction.

Introduction

Shop Customer Data is a detailed analysis of a imaginative shop's ideal customers. It helps a business to better understand its customers. The owner of a shop gets information about customers through membership cards. It is a popular dataset that provides valuable insights into the behaviour and preferences of customers who visit a particular shop. This collection of data contains various features such as age, gender, occupation, annual income, and spending score. This dataset is widely used by businesses to develop effective marketing strategies, improve customer satisfaction, and increase sales. The data in this dataset was collected through a survey conducted at the shop and is representative of a diverse range of customers in terms of age, gender, and income. The dataset includes 200 records, with each record containing five features: Customer ID, Gender, Age, Annual Income, and Spending Score (on a scale of 1-100).
The Shop Customer Dataset is widely used in data analysis, data mining, and machine learning applications and by analysing this dataset we can understand that which products are popular among certain demographics or what factors influence customers' purchasing decisions.

Implementation

Tool used: Orange

Orange (3.31.0) is an open-source data mining and visualization toolkit. It is used for explorative rapid qualitative analysis and interactive data visualization.

Data Description

The dataset about the customers from a shop was obtained from Kaggle website. The details about the attributes provided here are used for classifying the customers and find the spending score of each customers by evaluating their income and other features. This dataset contains 2000 instances and 8 features out of which 6 are numeric values and other 2 are categorical values. The attributes in the dataset are:

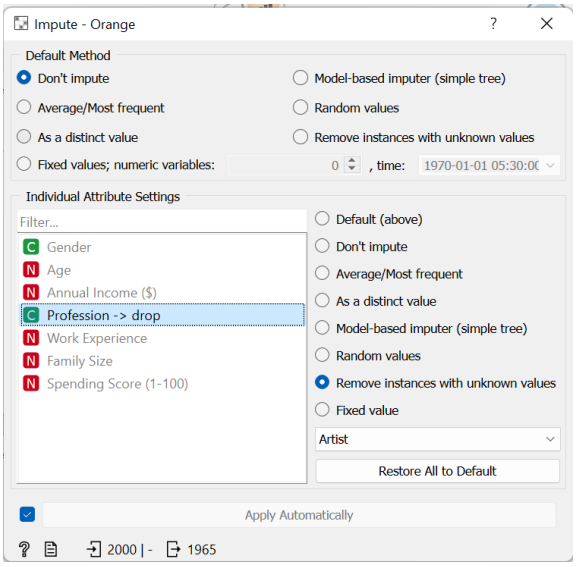
- CustomerID
- Gender
- Age
- Profession
- Work Experience
- Family Size
- Annual Income
- Spending Score (1-100)

No	Attribute	Description	Type
1.	CustomerID	This attribute contains the ID of each customer.	Numeric
2.	Gender	This feature shows the gender of the customer which takes two values like male or female.	Categorical
3.	Age	This shows the gender of the customer; either male or female.	Numeric
4.	Profession	The profession of each customer is being recorded. The professions included are engineer, lawyer, entertainer, artist, doctor, executive,	Categorical

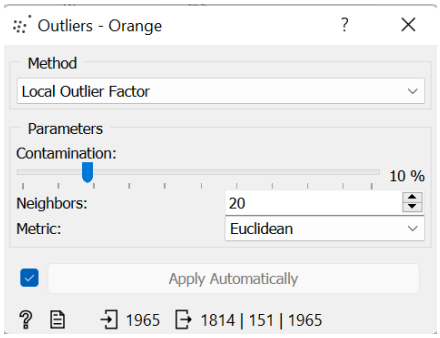
		homemaker, healthcare and marketing.	
5.	Work Experience	This feature shows the experience of the customer in their profession. The value ranges from 1 to 17.	Numeric
6.	Family Size	The values in this column depicts the number of dependents the customer has. This feature thus helps in analysing the financial status of the customer. The values present in this column ranges from 1 to 9.	Numeric
7.	Spending Score	This feature depicts the spending rate of the customer by analysing the income and other such attributes. Here the score is set between 1 and 100. This feature is set as the target variable.	Numeric

Data Pre-processing

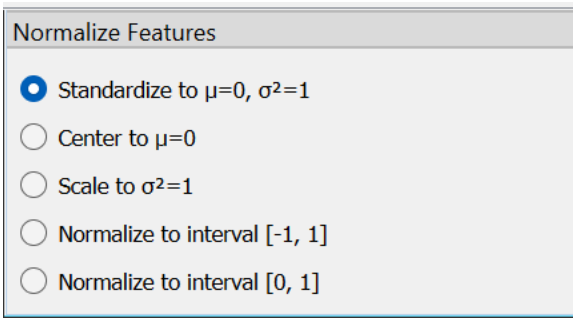
The dataset contains certain missing values which are to be cleared. Initially we remove the unwanted columns which are not required for the model building. At first, we removed the attribute CustomerID. Imputer is the widget which is used to impute the missing values or the null values. Such values were imputed by removing those values from the dataset since we have only 35 missing data in Profession attribute. Removing few datas will not affect the dataset.



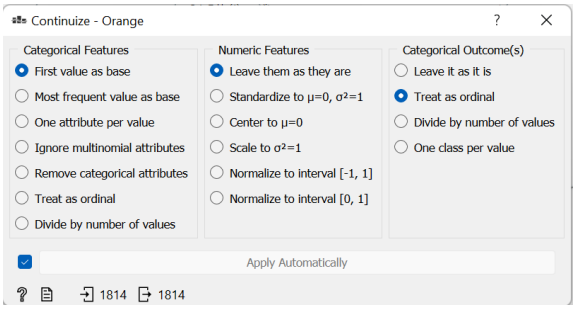
After removing the values, Box Plot is plotted in order to find the outliers of the data. Those data are removed using the widget Outliers.



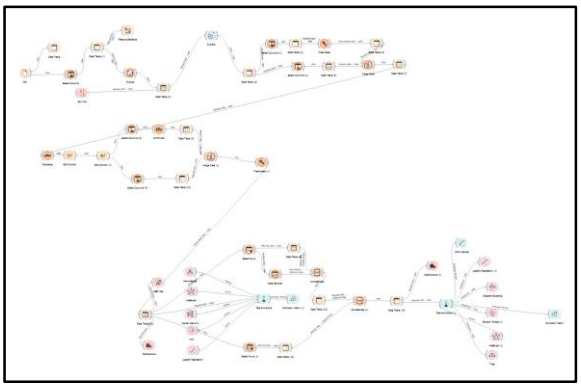
Then we go for further pre-processing. The annual income of the customer is made to categorical for building an accurate model for the dataset. For that, the attribute alone is chosen and the income values are normalized. Normalization is a technique used in data pre-processing to transform the values of a dataset to a common scale. This is done to improve the performance and accuracy of machine learning algorithms that are sensitive to the scale of the input data. It is done to decrease the range between the values in that dataset and align the values within a defined or fixed range.



Later, these normalized values are merged along with the rest of the attributes and made into complete data. Here, the target variable contains multiple values which will affect the effective model building. Hence this multiple values are discretized into two ranges, that is, below 50 and above 50. This attribute contains values till 100 since it is the spending score of the customers. The discretized data is then given values 0 and 1 for those data below 50 and data above 50 respectively.



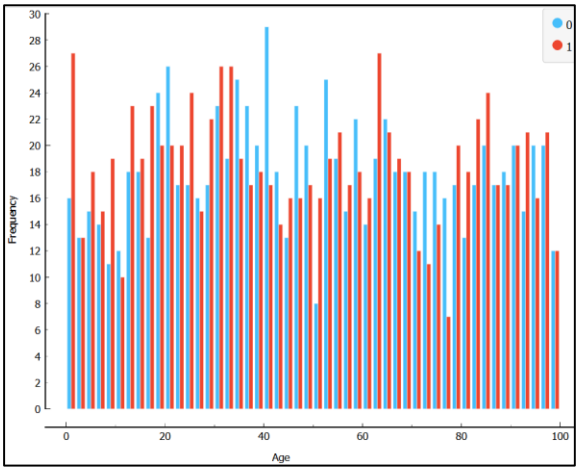
After discretizing the required values, the Profession column alone is continuized to find the profession of the customers. The entire data is then merged together to do further processes. These are the pre-processing which was done in this dataset. This pre-processed data can be used for visualization.



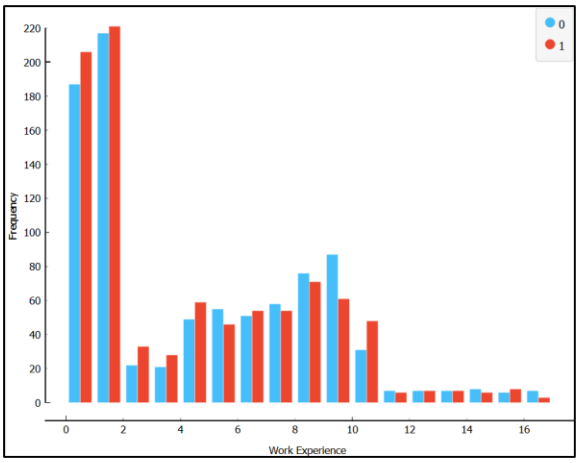
This is the data diagram done in Orange. The entire process of pre-processing and model building is done in this data diagram.

Data Visualization

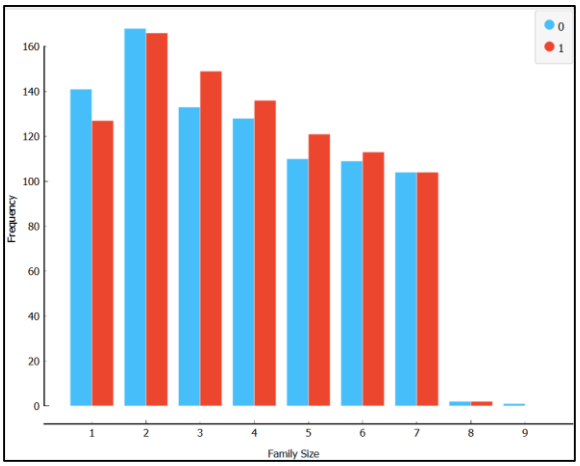
In order to analyse data and bring out any information, we need to explore the data. The graph below shows the distribution of age plotted along spending score. Here, the age extends up to 80-85. The red colour which represents the value 1 implies that those are the people with a spending score above 50 while the blue colour with value 0 implies the people who has a spending score below 50.



From this graph, we can see that people between age group of 20-40 spend more comparatively with other age groups. The old age group also has a high spending score.



This shows that customers with a work experience of 0-2 years has more spending score. Those with an experience range up to 10 years also have a significant spending score. Those below 12 years doesn't spend much because those people will be old aged.



This depicts that customers with dependent people 0-2 has a higher spending score compared with those customers having more than 2 dependents.

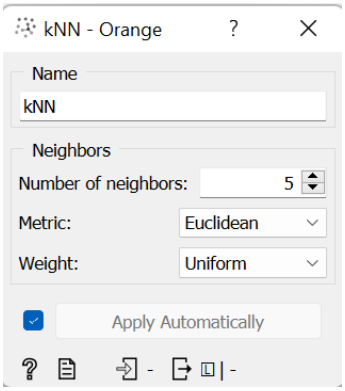
From these visualization we can say that spending of the youth as well as those with less number of dependents are3 high which means greater than the score of 50.

Model Building

After doing the entire pre-processing, models were made and accuracy of different models were evaluated.

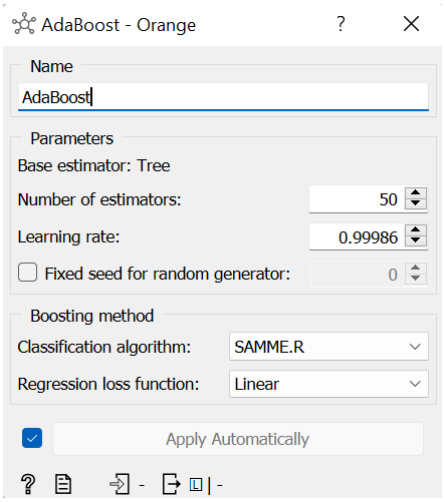
1. K-Nearest Neighbour Algorithm (KNN)

It is a type of supervised machine learning algorithm used for classification and regression tasks. In KNN, the algorithm predicts the class of a new data point based on the class of its k nearest neighbours in the training set. The class of the new data point is then assigned based on the majority class among its k nearest neighbours.



2. AdaBoost

AdaBoost (Adaptive Boosting) is a machine learning algorithm used for classification and regression tasks. It is a type of ensemble learning algorithm that combines multiple weak classifiers to create a strong classifier. The algorithm works by iteratively training a sequence of weak classifiers on the data, where each weak classifier is trained on a weighted version of the training set. After each iteration, the weights of the misclassified examples are increased so that subsequent classifiers focus on the examples that were previously misclassified.



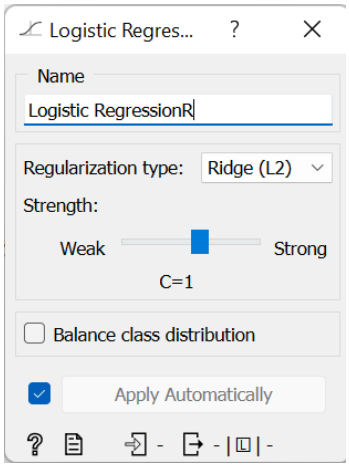
3. Naïve Bayes Algorithm

Naive Bayes algorithm is a type of supervised machine learning algorithm used for classification tasks. It is based on the Bayes' theorem and the assumption of conditional independence between the features. The algorithm works by first computing the prior probabilities of each class based on the frequency of class labels in the training data. Then, given a new data point, the algorithm computes the conditional probability of each class given the features of the data point using Bayes' theorem. Finally, the algorithm predicts the class label of the new data point as the class with the highest posterior probability.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

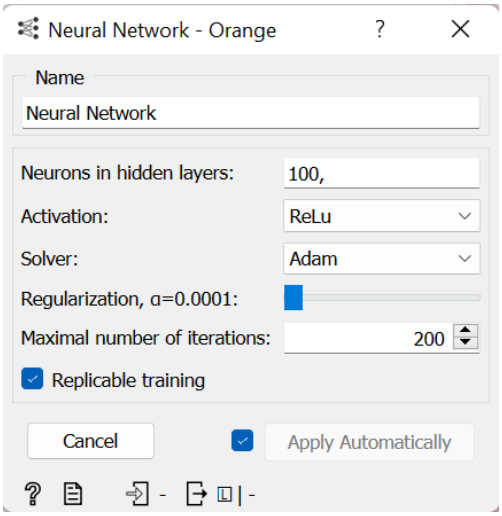
4. Logistic Regression

Logistic regression is a type of supervised machine learning algorithm used for binary classification tasks. It models the probability of the outcome the binary class label as a function of the features. The algorithm works by first fitting a linear regression model to the data, which estimates the relationship between the input variables and the continuous outcome variable.



5. Neural Network

A neural network is a type of machine learning algorithm inspired by the structure and function of the human brain. It consists of interconnected nodes, called neurons, that work together to learn patterns in data. The network is typically organized into layers, with each layer performing a different transformation on the input data. The input layer receives the raw data, and the output layer produces the final predictions or classifications.



All these models were added to the dataset and the accuracy, error and are under curve were evaluated.

Evaluation results for target 1

Model	AUC	CA	F1	Precision	Recall
AdaBoost	0.499	0.499	0.506	0.505	0.507
kNN	0.486	0.494	0.511	0.501	0.523
Logistic Regression	0.528	0.524	0.587	0.523	0.669
Naive Bayes	0.514	0.512	0.547	0.516	0.583
Neural Network	0.512	0.508	0.538	0.513	0.566

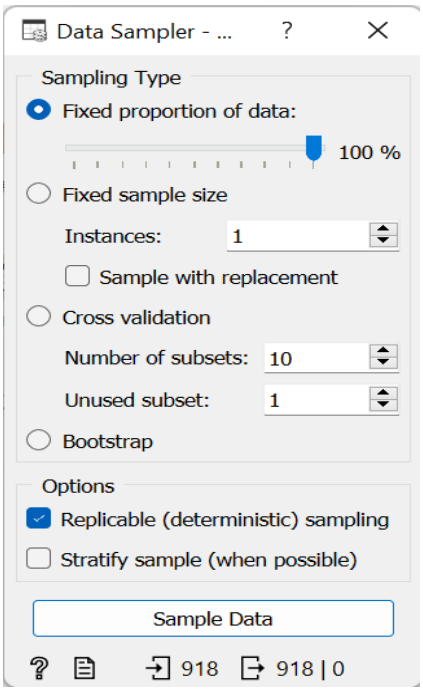
This is the accuracy, error, recall and precision of the data models obtained when the evaluation results for target is set as 1.

Evaluation results for target 0

Model	AUC	CA	F1	Precision	Recall
AdaBoost	0.499	0.499	0.493	0.493	0.492
kNN	0.486	0.494	0.476	0.488	0.465
Logistic Regression	0.528	0.524	0.438	0.525	0.375
Naive Bayes	0.514	0.512	0.471	0.507	0.440
Neural Network	0.512	0.508	0.474	0.502	0.449

This is the accuracy, error, recall and precision of the data models obtained when the evaluation results for target is set as 0.

From both these data, we could understand that the value for precision and recall for the models is higher for the target variable for 1. Hence we divide the data based on the both the values in the target variable and then determining models for the dataset. Data with target variable 1 is separated and the data is sampled with a proportion of 100% and then it is concatenated with the data having target variable 0.



A better data model with a better accuracy, precision and recall can be obtained by focusing on the value giving higher accuracy.

Results

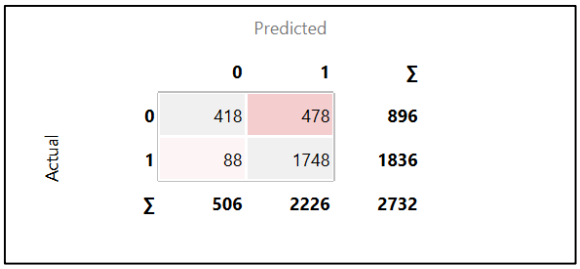
The model obtained after sampling the wanted data gives better accuracy and less errors. The target variable is the spending score which has been classified into 0 and 1. The models deployed to the

processed datasets are logistic regression, geadient boosting, AdaBoost, tree and random forest. Among this, Adaboost gives the maximum accuracy of 79%.

Evaluation results for target 1

Model	AUC	CA	F1	Precision	Recall
AdaBoost (1)	0.709	0.793	0.861	0.785	0.952

A confusion matrix is a table that is often used to evaluate the performance of a classification model by comparing the actual values with the predicted values. In the context of a shop customer dataset, a confusion matrix could be used to evaluate the performance of a model that predicts whether or not a customer is likely to make a purchase based on their demographic and behavioral characteristics. The confusion matrix of the AdaBoost is given below:



All the features are required for this effective model building. But the main features which contribute more are annual income, family members, profession and work experience.

Inferences and Conclusions

The main goal of this dataset is to analyse and predict the spending of an individual or a customer. An individual can spent money according to many factors affecting him. Few of the factors were mentioned in this dataset collected from a shop. The features included were annual income, family dependents, profession, work experience, age and gender. The spending score which ranges from 1-100 is also recorded of each customer. This spending score is determined based on the purchases made by the customer from that particular shop. These data have been pre-processed and made into structured data using different pre-processing methods and models are deployed to the dataset. From the analysis done, we can say that

i) the one with few years of work experience and with maximum of 5 members spends more money compared to the other category of people

ii) ii) customers spent money irrespective of their ages

iii) iii) people with highest salary spend more money when compared to the people with low salary.

With the help of these conclusions, if an individual's annual income, family size and work experience are given we can predict the spending score of that individual. Among the applied classification models, AdaBoost gives the highest accuracy with minimum error for classifying customer into their spending score.

References

1. https://www.kaggle.com/code/svinothas/analyzing-customer-data
2. https://www.wikipedia.org/