Title: Perspectives of Data Science Project

Objective: Implementing a classification model on the given data set to predict and analyze the gender of the person using selected features such as Height, weight, age, marital status and no of shopping trips in a month.

Introduction:

As we decided to predict and analyze the gender of the person using features we selected from the data set, first we need to clean the data by using data cleaning method where we identify the incorrect, incomplete and in accurate data and then by modifying, replacing or deleting them according to the necessity. Next, we do the EDA (Exploratory Data Analysis) to discover patterns by using statistical summary and graphical representations. And then we used different methods available to get better output for our analysis part.

Summary:

Data Cleaning is the process of identifying the incorrect, incomplete, inaccurate, irrelevant or missing part of the data and then modifying, replacing or deleting them according the necessity. We performed data cleaning by removing all the null values, created dummies for gender and marital status. Next, we have formatted everything from the selected features into the respective units and type conversions.

Exploratory Data Analysis is an approach of analyzing the data sets to summarize the data sets by their characteristics using different statistical methods and data visualization process.

Feature selection is simple yet effective way to eliminate redundant and irrelevant data which improves accuracy, reduces the computation time and also facilitates to understand the data more in depth.

There are three types of feature selection methods: 1.Wrapper methods(Forward, Backward, stepwise)

2.Filter methods (Anova, pearson correlation, variance thresholding)

3. Embeded Methods(Lasso, Ridge, Decision Tree)

From all the three metods we have used embedded method(Lasso) as we expect it is better and more usefull in finding out the output.  
Lasso Regression is a regression which allows us to shrink the data or regularize the coefficients to avoid over fitting and make them work better on different data sets.By performing the lasso regression out of 5 features we selected it deducted only one and settles with other four features.

ABSTRACT: Implementing a classification model on the given data set to predict and analyze the gender of the person using selected features such as Height, weight, age, marital status and no of shopping trips in a month.

Introduction:

Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets. Data science takes up challenges such as cleaning, capturing and transforming the unstructured data using different technologies to store and process. The next stage involves data processing which creates effective data by mining, classification, clustering, modelling and summarizing. Next comes the data analysis part where we conduct exploratory work, regression, predictive analysis and qualitative analysis. And now we visualize the data which involves data visualization, data reporting and smarter decision making. . Data science applications are frequently used in healthcare, marketing, banking and finance.

In the present world, the generation and application of information is a critical economic activity.

Data is a precious asset of any organization. It helps firms understand and enhance their processes, thereby saving time and money. Wastage of time and money, such as a terrible advertising decision, can deplete resources and severely impact a business. The efficient use of data enables businesses to reduce such wastage

Data is meaningless until its conversion into valuable information.

According to IDC, by 2025, global data will grow to 175 zettabytes.

The organizational importance of Data Science is continuously increasing. According to one study, the global Data Science market is expected to grow to $115 billion by 2023.

 Data Science enables companies to efficiently understand gigantic data from multiple sources and derive valuable insights to make smarter data-driven decisions.

As we decided to predict and analyze the gender of the person using features we selected from the data set, first we need to clean the data by using data cleaning method where we identify the incorrect, incomplete and in accurate data and then by modifying, replacing or deleting them according to the necessity. Next, we do the EDA (Exploratory Data Analysis) to discover patterns by using statistical summary and graphical representations. And then we used different methods available to get better output for our analysis part.

Logistic Regression is a statical analysis method used to predict the outcome based on prior observations of the data set. This method predicts a dependent variable by analyzing the relationship between one or more existing independent variables. We preferred because of the l;ogistic regression is in binary format.

A [**confusion matrix**](https://www.educative.io/edpresso/what-is-a-confusion-matrix) is a tabular summary of the number of correct and incorrect predictions made by a classifier. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.

Here by using the logistic regression model we divided the data into train and split cases and by fitting the data into model we got the score of model 0.92 and accuracy of the model 0.88.

The confusion matrix in case 0 it is predicting 3 correct and 2 wrong out of 5 where as in case 1 it predicts 19 correct and 1 wrong out of 20. From the above we can say it is inbalanced data as it is showing 5 girls and 20 boys.

AUC is an **area under ROC curve**. The value of AUC characterizes the model performance. Higher the AUC value, higher the performance of the model. The perfect classifier will have high value of true positive rate and low value of false positive rate.

By performing AUC of the logistic regression we got 0.91, we can understand that there is high chance that the classifier will be able to distinguish the positive class values from the negative class values.

Lasso regression is a type of linear regression that allows you to shrink or regularize the co effiecients to avaoid overfitting and making them work better on different data sets showing high multicollinearity or we can do variable elimination and feature selection. The main goal is to obtain the subset of predictors that minimizes the prediction error for a quantitative response variable.

We used LaasoCv to pick the best features, so laaso picked 3 variables(Height, Weight, Number of shopping trips) and eliminated the other 2 variables(Age, Marital Status)Assigning the ‘X’variable to the Height, Weight, No. of shopping trips removing the lasso eliminated variables, the score of the logistic regression after eliminating variables is 0.90

By predicting using lasso regression the accuracy of the model is 0.86 and the confusion matrix in case 0 it predicts 6 correct and 2 wrong out of 8 and in case 1 it predicts 20 correct and 2 wrong out of 22.

Now the data is imbalanced as we can see the accuracy for the logistic regression is 0.88 and for lasso regression is 0.86, so we are going to do Balance the data by using the Random Sampling method.

After performing the code required for balancing the data we got the resample data set shape as (144,3) where male and female are of 72 in number.

Now performing the logistic regression for the balanced data the score of the model is 00.91 and accuracy of the model os 0.84 only and the confusion matrix in case 0 it predicts 24 correct and 2 wrong out of 24 and in case 1 it is predicting the 13 correct and 5 wrong out of 18.