**PREDICTIVE MODELLING**

Name: Vamsi Krishna Bommu

Batch: PGPDSBA.APRIL23.B

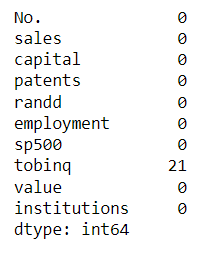
**CONTENTS**

* Problem 1
* EDA
* Univariate Analysis
* Bivariate Analysis
* Model Creation
* Inference: Business Insights and Recommendations
* Problem 2
* Data Ingestion and Exploratory Data Analysis
* Descriptive Statistics
* Univariate Analysis
* Bivariate Analysis
* Data Encoding
* Model Creation
* Inference and Recommendation

**Problem 1: Linear Regression You are a part of an investment firm, and your work is to do research about these 759 firms. You are provided with the dataset containing the sales and other attributes of these 759 firms. Predict the sales of these firms on the basis of the details given in the dataset so as to help your company in investing consciously. Also, provide them with 5 attributes that are most important.**

**Exploratory Data Analysis:**

* The dataset includes various attributes for 759 firms and having sales as the target variable.
* Checking the null values across all the columns presented in the dataset and getting the type of the values presented in each column for eg., Dtype: int64. There were 21 null values in tobinq columns.



* Getting the shape of the dataset using data.shape. It consists of 759 rows and 10 columns. And different types of data types were presented in the dataset.

A screenshot of a computer

Description automatically generated

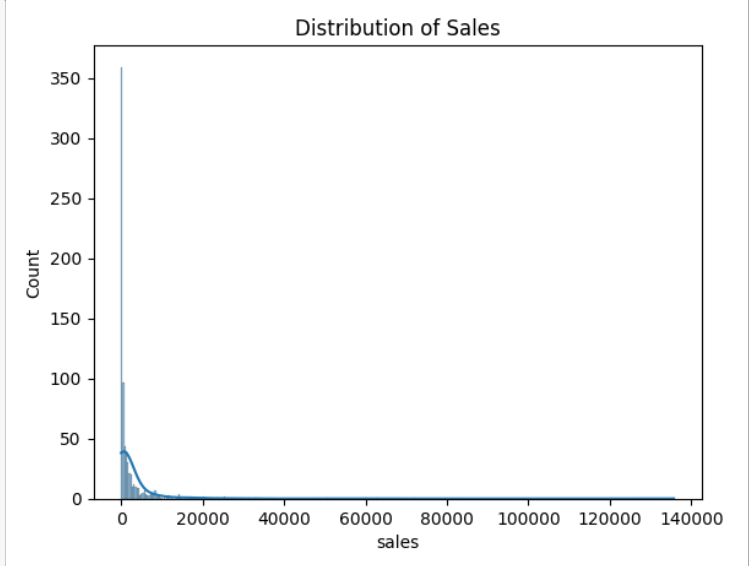
* To view the data in the descriptive manner, we use data.describe() method to get the statistical summary of numerical columns, offering insights into mean, median, standard deviation, and distribution range of each numeric attribute. This is important for identifying outliers, understanding data spread, and preliminary insights into data normalization needs.

A screenshot of a computer screen

Description automatically generated

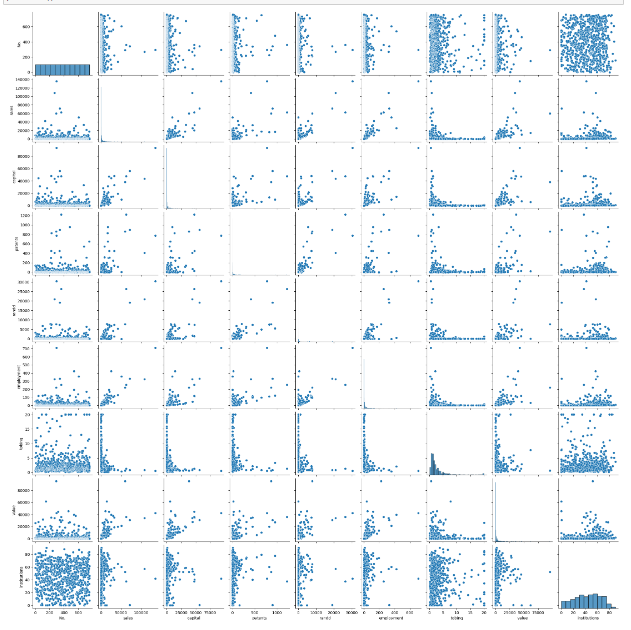
**Univariate Analysis:**

* A histogram with a kernel density estimate (KDE) for the 'sales' column was plotted, showing the distribution of sales across the firms. This visualization is essential for understanding the distribution's skewness, detecting outliers, and determining the sales performance of firms briefly.



**Bivariate Analysis:**

* Pairwise relationships between all variables were visualized using sns.pairplot(data), providing insights into the relationships and correlations between sales and other attributes. This is crucial for identifying potential predictors for the regression model.



**Handling Null Values:**

* Null values were identified and filled with the mean of their respective columns. This approach is used for the continuous numerical variables. However, the approach might vary depending on the variable’s nature and distribution.
* The necessity for scaling was not directly addressed. Given that linear regression can be sensitive to the scale of input features, especially when interpreting coefficients, scaling could be beneficial for models involving variables with vastly different scales and units. However, the direct impact of scaling on model performance (R-square, RMSE) was not assessed.

**Data Handling:**

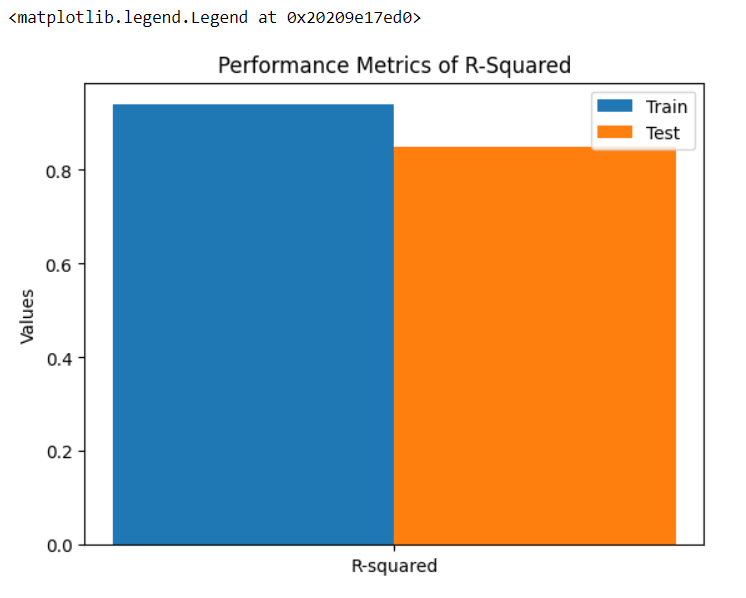
* Non-numeric values were identified, and a binary encoding was applied to replace 'yes' and 'no' with 1 and 0, respectively. This encoding is suitable for binary categorical variables but might not be adequate for nominal variables with more than two categories if present.

**Splitting the Dataset:**

* The data was split into a training set (70%) and a testing set (30%), following common practices for model validation.

**Model Creation:**

* A linear regression model was applied and trained, and its performance was evaluated using R-square and RMSE on both training and testing sets.

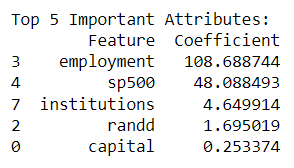


A screen shot of a graph

Description automatically generated

**Inference: Business Insights and Recommendations:**

* The model demonstrated reasonable predictive ability, with R-square values indicating a decent level of variance explanation in both training and testing sets. However, a slight overfitting was noted, as indicated by a higher R-square on the training set.
* RMSE values provided insights into the average prediction error, essential for assessing model accuracy and reliability.
* Top 5 Attributes: Identifying the five most important attributes based on their coefficients suggests areas where the firm could focus to improve sales or areas of due diligence for investment decisions.



**Problem 2: Logistic Regression and Linear Discriminant Analysis. You are hired by the Government to do an analysis of car crashes. You are provided details of car crashes, among which some people survived and some didn't. You have to help the government in predicting whether a person will survive or not on the basis of the information given in the data set so as to provide insights that will help the government to make stronger laws for car manufacturers to ensure safety measures. Also, find out the important factors on the basis of which you made your predictions.**

**Data Ingestion and Exploratory Data Analysis:**

* The dataset includes a range of variables associated with car crash incidents.

A screenshot of a computer

Description automatically generated

* Describing the dataset using df.describe() to provide the insights into the basic statistics and the structure of the data.

A table with numbers and letters

Description automatically generated

**Check Null Values:**

* Check the null values using isnull().sum(). The null value check indicated the presence and absence of missing values across all the features, and filling or removing those values to make the dataset clean for further process. In this dataset there is 77 null values presented in the injSeverity column.

A white background with black text

Description automatically generated

**Descriptive Statistics:**

* The descriptive statistics revealed the central tendency, dispersion, and shape of the dataset’s numerical features.
* This dataset acquires 11217 rows and 16 columns.

A screenshot of a computer

Description automatically generated

**Inference:**

* If any null values are presented in the data, then, it must be addressed appropriately to ensure the integrity of the modeling process. The choice of imputation could be based on the nature of missing values and the distribution of the data.

**Univariate Analysis:**

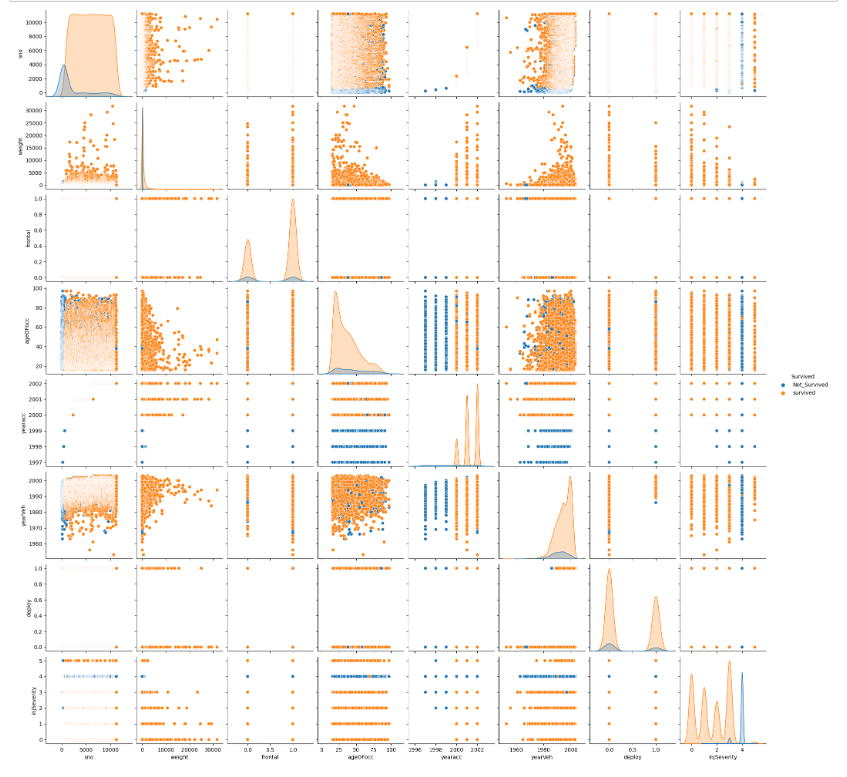
* In Univariate Analysis, we plotted histograms and count plots provided insights into the distribution of individual variables such as ‘age0Focc’, ‘dvcat’, ‘sex’, ‘airbag’, and ‘seatbelt’. This helped in understanding the skewness, kurtosis, and categorical balance in the dataset.

A group of graphs with different colored squares

Description automatically generated with medium confidence

**Bivariate Analysis:**

* The pairplot with hue=’Survived’ allowed us to observe the relationships between pairs of variables and how they correlate with the survival outcome.



**Data Encoding and Splitting of the Data:**

* Categorical variables were encoded into numerical formats using ‘Label Encoder’, facilitating their use in machine learning models.
* Splitting the data into 70 and 30 percent for training and testing purposes. This is performed using train\_test\_split method from sklearn library.
* This splitting method ensures a balanced approach for training and validation.

**Model Creation:**

* Both Logistic Regression and Linear Discriminant Analysis were applied to the dataset after handling missing values through mean imputation.
* Logistic Regression and LDA performance were evaluated based on Accuracy, Confusion Matrix, ROC Curve, and ROC\_AUC score.
* Both models exhibited comparable accuracy levels on training and test datasets, with the confusion matrix providing insights into true positives, false positives, true negatives, and false negatives.
* ROC Curve and ROC\_AUC score further helped in assessing the models' ability to distinguish between the classes.

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

**Inference and Recommendation:**

* Safety features such as airbag and seatbelt usage significantly impact survival chances in car crashes.
* The dvcat variable, representing speed categories, shows a strong correlation with survival outcomes, suggesting higher speeds are associated with lower survival rates.
* Enhancing vehicle safety features and promoting their use can significantly improve survival rates.
* Public awareness campaigns highlighting the dangers of high-speed driving could be beneficial.
* Further research could explore more sophisticated models or feature engineering to improve prediction accuracy.