ANALYSIS ON DONOR DATASET FOR NON-PROFIT ORGANIZATION

BY
HETAL GADA

APPROACH TO THE PROBLEM

- The goal of the analysis is to identify the factors that influence a donor's lifetime giving to a particular organization using a dataset containing demographic information and giving history of donors.

- Exploratory data analysis (EDA) was performed to gain insights into the distribution of the data and the relationships between variables

- Four different models (linear regression, decision tree, random forest, and SVM) were trained and evaluated to predict lifetime giving.

- R-squared values were used as a metric to compare the performance of the different models. The random forest model achieved the highest R-squared value, indicating that it was the best model for predicting lifetime giving.

THOUGHT PROCESS-1

- Started by exploring the dataset and understanding the variables to identify the target variable for the analysis
- Noticed that the dataset contained some missing values and decided to drop rows with over 50% missing values after analyzing the target variable.
- Applied label encoding to five categorical columns by grouping similar categories and assigning a unique number to each group.
- Conducted data cleaning by removing irrelevant columns that are not needed in the analysis of the target variable, which is total lifetime giving.

THOUGHT PROCESS-2

- Performed exploratory data analysis (EDA) to gain insights into the distribution of the data, relationships between variables, and identified outliers in the dataset.
- Identified the most optimal variables for predicting total lifetime giving: Using Lasso Regression by performing regularization
- Worked with different models like linear regression, decision tree, random forest, and SVM, and used R-squared values to compare their performance in predicting lifetime giving
- Found that the random forest model had the highest R-squared value and was the best model for predicting lifetime giving

METHODOLOGY

01

LOAD DATASET

Clean the dataset to ensure that there are no discrepancies

02

FEATURE ENGINEERING

Perform Exploratory Data Analysis (EDA) to know the data

Transformations, Feature Extractions

Feature Selection

03

MODEL CREATING AND TRAINING

Create the model by fitting the data using appropriate parameters

Train and Validate the Output

04

EVALUATION AND PREDICTION

Evaluate the model against previously unseen data

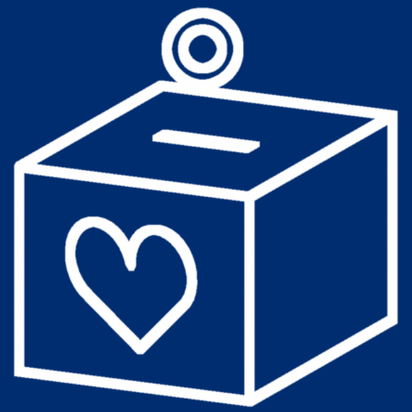
Tune model parameters for improved performance and test the model

LOAD DATASET

Dataset preparation is the most important aspect of creating efficient machine learning algorithms

- Check Data Type
- Missing Data
- Outliers
- Data Sampling

REGRESSION MODELS



- LINEAR REGRESSION
- RIDGE REGRESSION
- LASSO REGRESSION
- DECISION TREE
- RANDOM FOREST
- SUPPORT VECTOR MACHINE

```{r}
#Load the data:

df2 <- read_excel("/Users/hetal/Downloads/RedactedClientConstituent_File.xlsx")
head(df2)
```

A tibble: 6 × 31

CnBio_ID <dbl>	First Gift Date <S3: POSIXct>	Last Gift Date <S3: POSIXct>	Largest Gift Date <S3: POSIXct>	CnBio_DateAdded <S3: POSIXct>	CnBio_DateChanged <S3: POSIXct>
200001488	<NA>	<NA>	<NA>	2019-10-14	2021-10-11
200001489	<NA>	<NA>	<NA>	2019-10-14	2021-10-11
200001490	<NA>	<NA>	<NA>	2019-10-14	2021-10-12
200001491	<NA>	<NA>	<NA>	2019-10-14	2021-10-12
200001492	<NA>	<NA>	<NA>	2019-10-14	2021-10-11
200001493	<NA>	<NA>	<NA>	2019-10-14	2021-10-25

6 rows | 1–6 of 31 columns

DESCRIPTION OF THE DATASET

NUMBER OF ROWS - 42287

NUMBER OF COLUMNS - 31

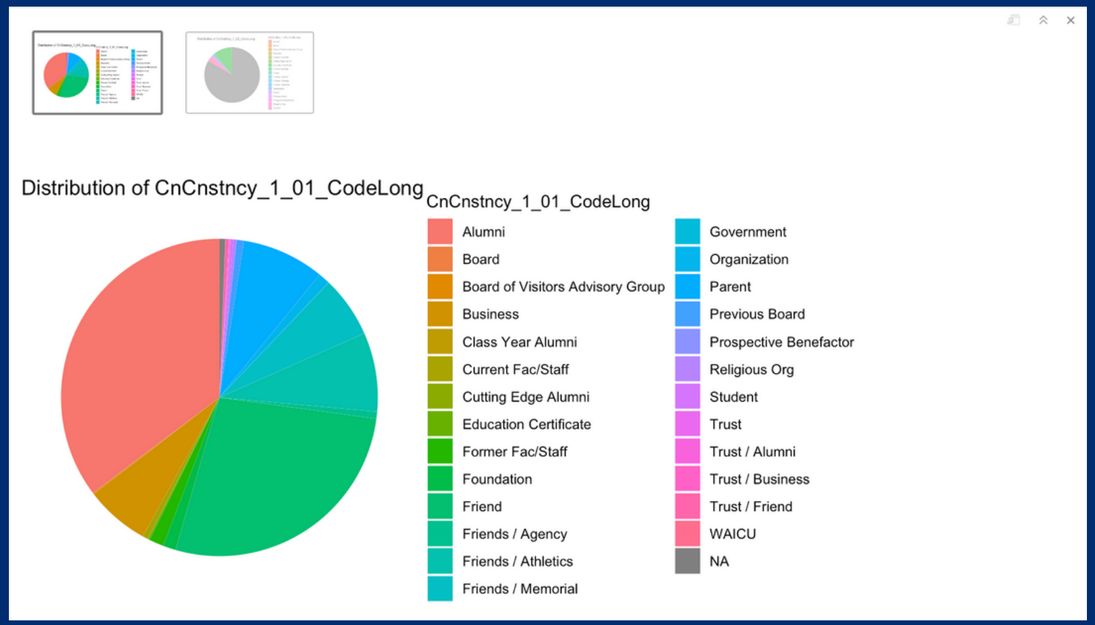
NUMBER OF NUMERICAL COLUMNS - 8

NUMBER OF CATEGORICAL COLUMNS - 18

NUMBER OF DATE COLUMNS - 5

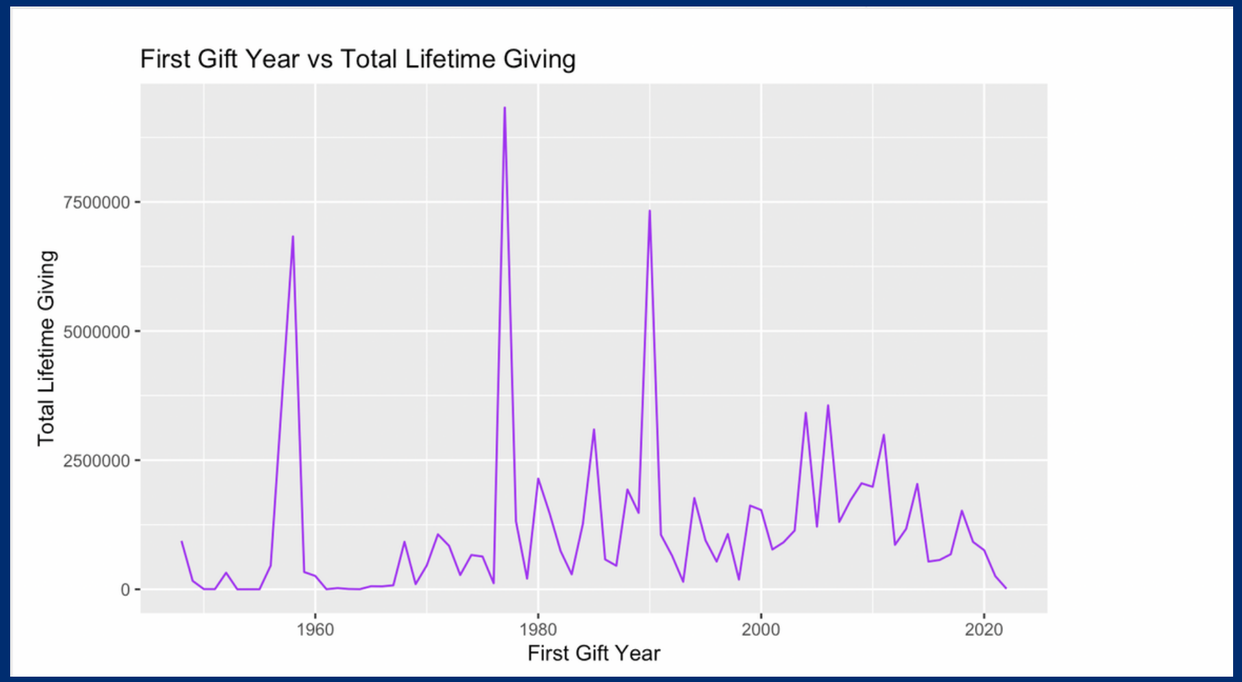
FORMAT - XLSX FILE

RESULT & ANALYSIS

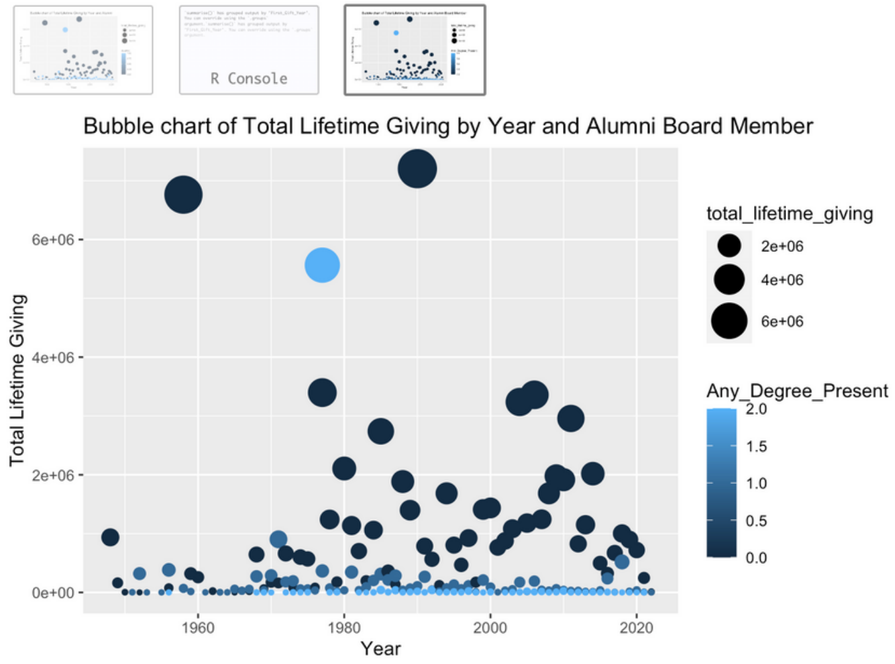


GROUPED THE SIMILAR CATEGORIES AND PERFORMED LABEL ENCODING

ANALYZED THE DATA FROM COLUMN "FIRST GIFT DATE" TO KNOW IF THEY HAVE INITIATED THE DONATED EVER BETWEEN: 1950 TO 2022

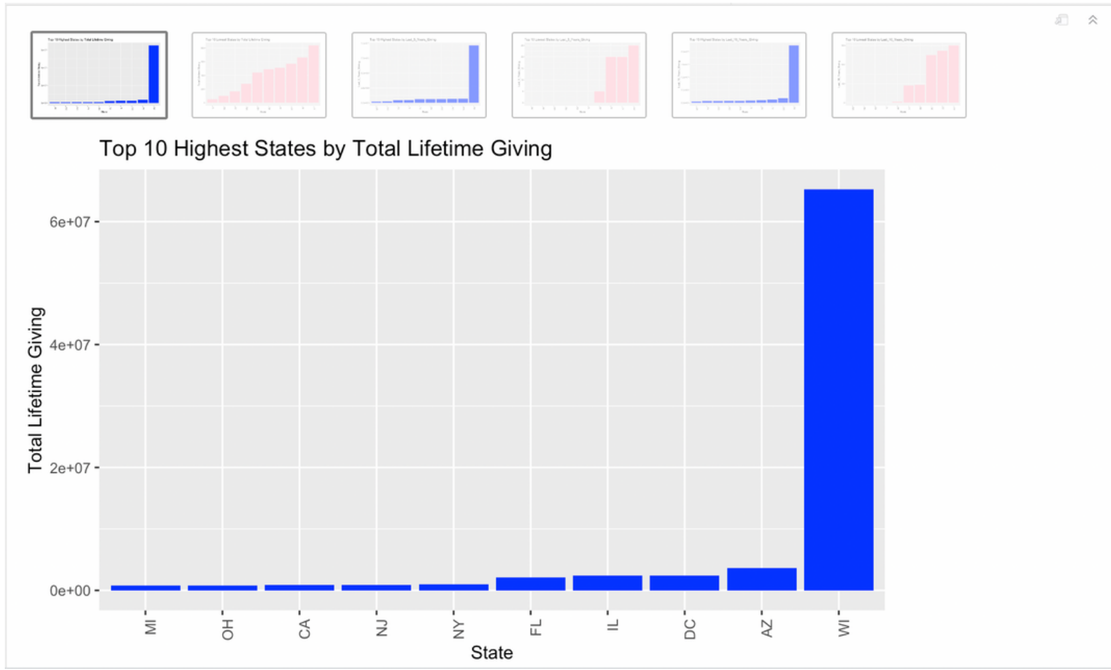


RESULT & ANALYSIS

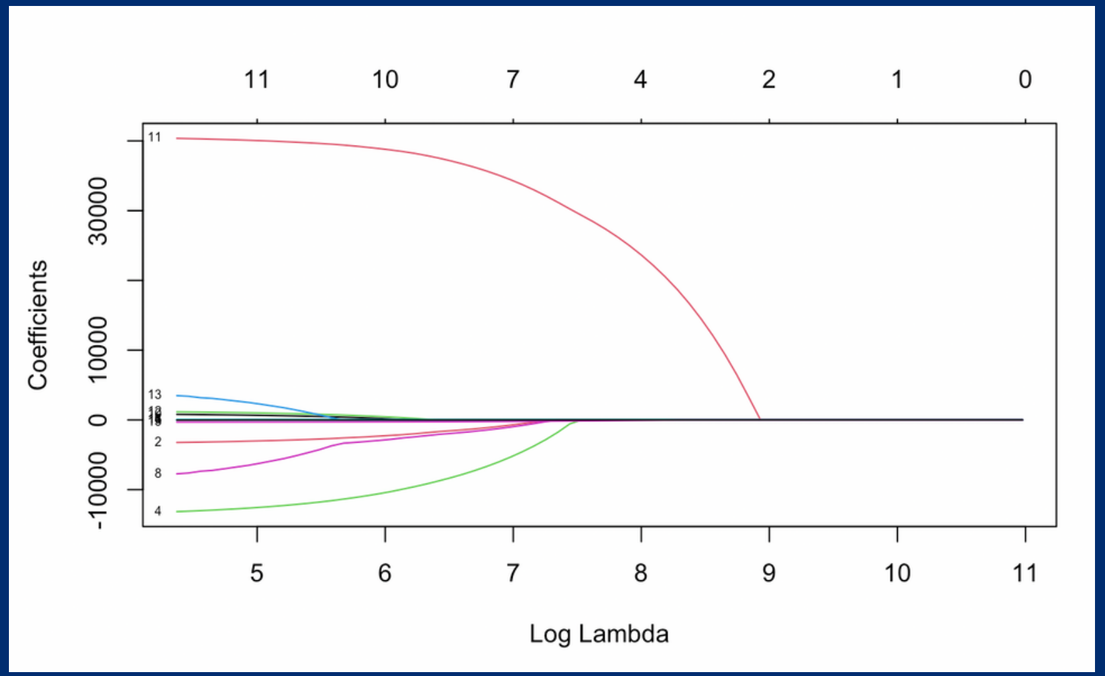


PERFORMED BUBBLE CHART TO ANALYZE ALUMNI BOARD MEMBER

PERFORMED ANALYSIS BASED ON STATES WITH ALL THE 3 GIVINGS AND CHECK IF THEY ARE CONSISTENT ON DONATION OR NOT (LAST 5 YEARS & 10 YEARS)

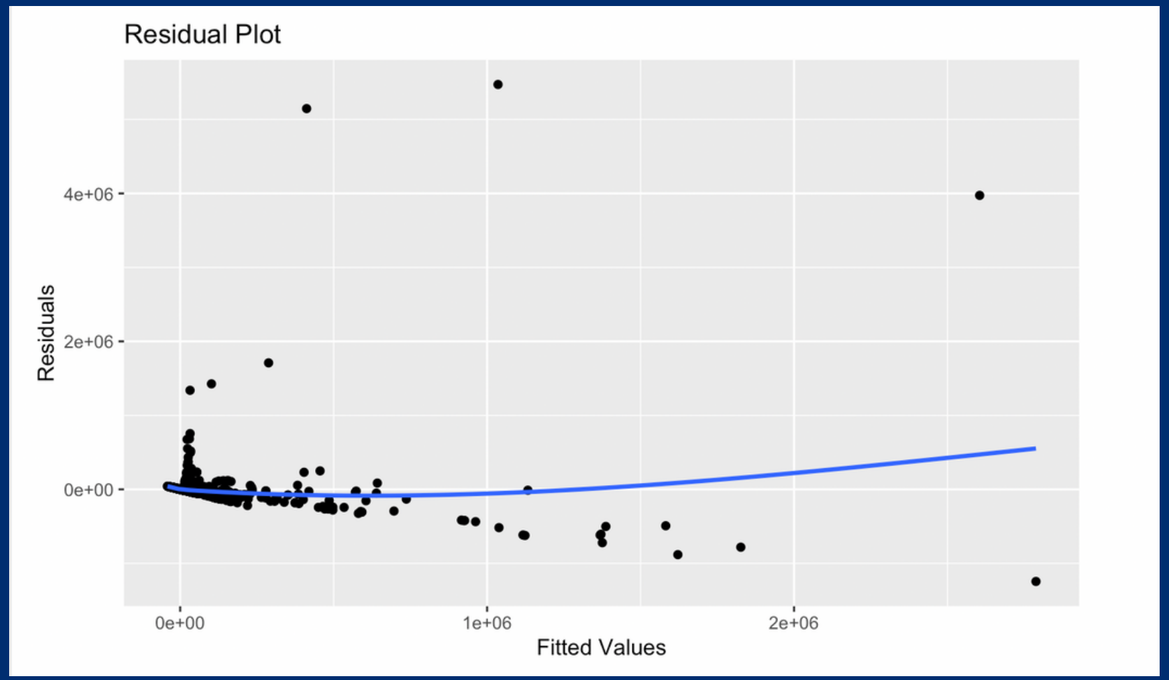


RESULT & ANALYSIS

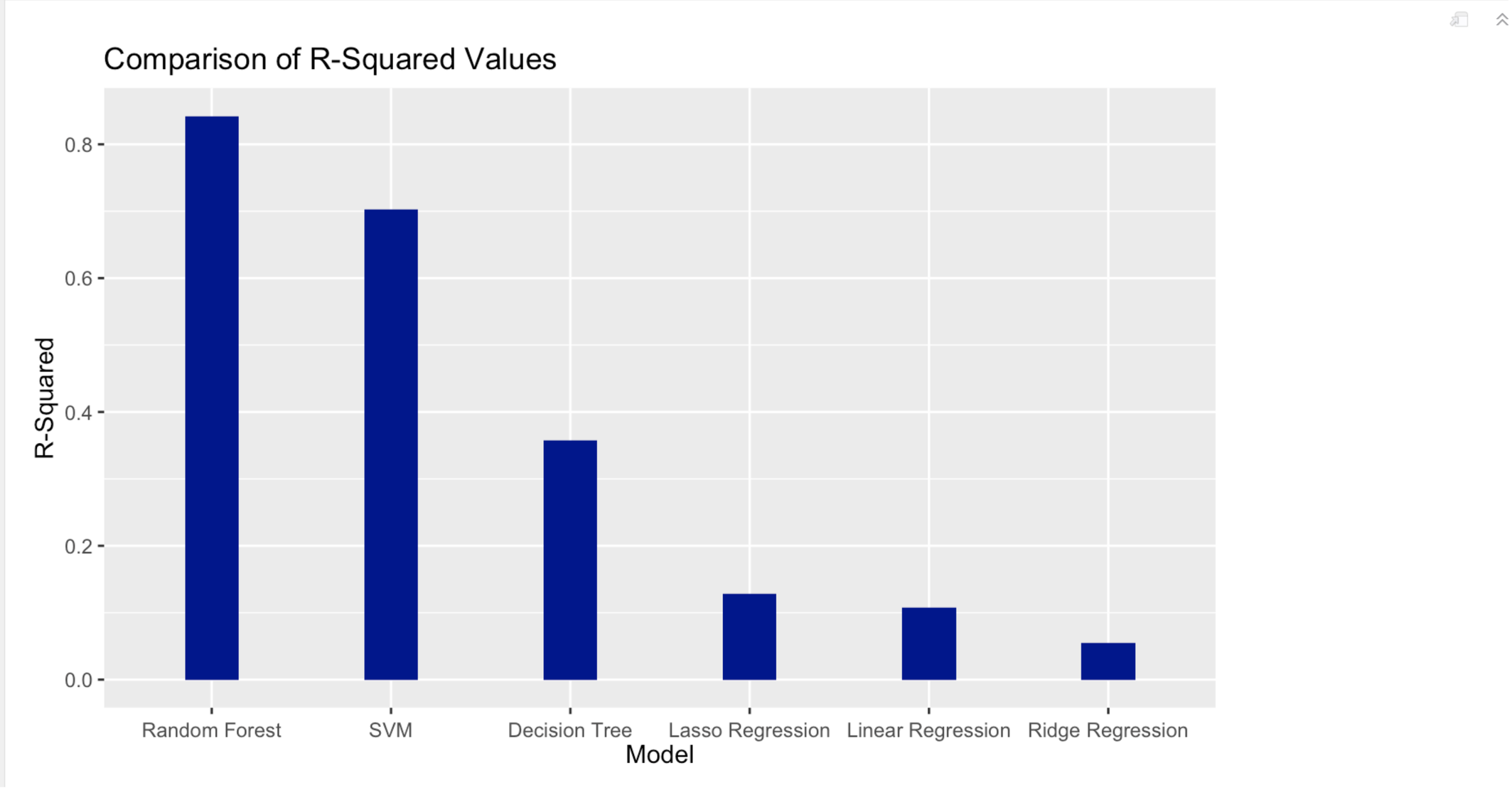


PERFORMED TO FIND THE OPTIMAL FEATURES USING LASSO REGRESSION

RESIDUAL V/S FITTED GRAPH



RESULT & ANALYSIS



Thank You !