# Loan Application Status Prediction

## Track Your Application Status

Answer all of the security questions below to get your Application Status. Please make sure that the information you provide below matches what we have on record for you.

| | |
|---|---|
| Application Id | |
| Mobile | XXXXXXXXXX |

If your information matches our records, you will be able to track your application.

**Submitted by:**

**Abhilekh Verma**

## Contents

# ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. I had also taken help from YouTube & online videos.

# INTRODUCTION

## PROBLEM STATEMENT

We have all been in situation where we go to a doctor in emergency and find that the consultation fees are too high. As a data scientist we all should do better. What if we have data that records important details about a doctor, and we get to build a model to predict the doctor’s consulting fee.? This is the use case that lets we do that.

## ABOUT DATASET

We have two datasets. Train data and Test data. We have to build a model on Train data and then after Predict that model on the Test data.

## DATA ANALYSIS

### COLUMNS IN THE DATAFRAME

- Qualification: Qualification and degrees held by the doctor
- Experience: Experience of the doctor in number of years
- Rating: Rating given by patients
- Profile: Type of the doctor
- Miscellaneous Info: Extra information about the doctor
- Place: Area and the city where the doctor is located.
- Fees: Fees charged by the doctor (Target Variable)

### OBJECTIVE

- Our main objective is to predict the fees of the doctors using machine learning algorithms.
- All the parameters will be analysed through Machine Learning algorithms like Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression etc which will help to predict the fees.

### DATA DESCRIPTION

- The dataset contains Qualification, Experience, Rating, Profile, Miscellaneous Info, Place, and Fees.
- The source of data is taken from GitHub. (https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Doctor_fee_consultation.zip)

### METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the sales because it gives the best outcome.
- The dependent variable is Fees, whereas other variables i.e., Qualification, Experience, Rating, Profile, Miscellaneous Info, Place is the independent variables.

### METRIC USAGE

a. Linear Regression.
b. Lasso Regression.
c. Ridge Regression.
d. Elastic Net Regression.
e. Decision Tree Regressor.
f. Random Forest Regressor.

# SYSTEM REQUIREMENTS

### Hardware and Software Requirements and Tools Used

a) Hardware Requirement:
   - i. Intel core i5
   - ii. 8 GB Ram

b) Software Requirement:
   - i. Python 3.x with packages:
     1. Pandas: Data analysis and manipulation tool
     2. NumPy: Provide support for mathematical functions, random number etc.
     3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
     4. Seaborn: is a library mostly used for statistical plotting in python.
     5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

## APPROACH

### IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- R2 score: is used to evaluate the performance of a linear regression mode.
- Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
- Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
- Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
- Root Mean Squared error : is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
- Z-score: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- Decision Tree Regression: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.
- Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
- Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting..

### TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Linear Regression
c. Lasso Regression
d. Ridge Regression
e. Elastic Net Regression
f. Grid Search CV
g. Cross Validation
h. Decision Tree Regressor
i. Random Forest Regressor

**KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION**

- Analysed data for any unique values.
- Analysed data for distribution.
- Checked outliers through z-score method.
- checked skewness present in the dataset.
- Done Standard Scaling.
- Cross validate the r2 score from overfitting.
- Hyper Parameter tuning using Grid Search CV

## CONCLUSION

After checking r2 score, cross validation, Ensemble Techniques, we came to the conclusion Linear Regression predicting 100% accuracy is best suited model for our purpose of predicting the Loan application status.

### LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs.  These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

## REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.