Northeastern University

Insurance Data Analysis using ML Techniques

Data Science Methods & Tools - Mid Term Project

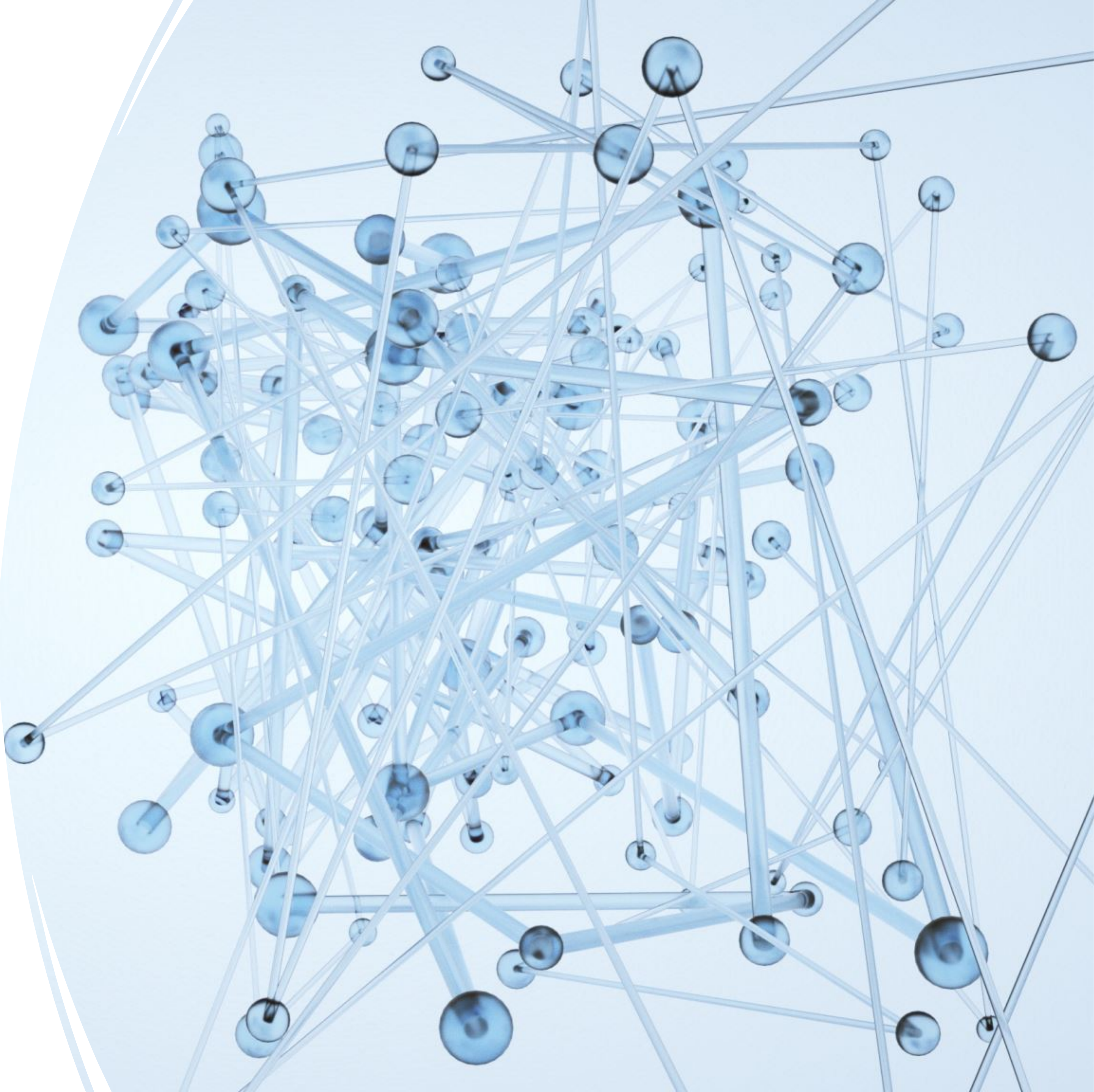
Presented By: Group 1

Anshita Verma: (001007320)

Mani Deepak Reddy:(002728148)

Uddhav Zambare: (002199488)

Revati Lachyan: (002197827)



Agenda

- Problem Statement
- Data Description
- Data Cleaning
- Data Preprocessing
- Machine Learning Approaches
- Accuracy
- Conclusion

Problem Statement:

Apply FOUR Machine Learning techniques on Prudential life insurance dataset and compare their performance.

Dataset:

Prudential Life Insurance dataset available on Kaggle.

Data Description

Prudential life insurance dataset with 'Ordinal' Ta rget variable which describes the risk values from 1-8 with **1** being '**Highest Risk**' and **8** being '**Lowest Risk**'.

Categorical

- Product_Info_1,2,3,5,6,7
- Employment_Info_2,3,5
- InsuredInfo_1,2,3,4,5,6,7
- Insurance_History_1,2,3,4,5,6,7,8,9
- Medical_History_2 - 41

Dummy Variable

- Medical_Keyword_1-48

Discrete

- Medical_History_1,10,15,24,32

Continuous

- Ins_Age, Ht, Wt, BMI,
- Product_Info_4,
- Employment_Info_1,4,6,
- Insurance_History_5,
- Family_Hist_2,3,4,5

Northeastern
University

Data Cleaning

- Handled Missing Values
- Removed Columns with Null Values over 30 %
- Performed Imputing
  - Mean
  - Median

Data Preprocessing

- I to C encoding
  - Converted categorical column to numerical.
- Dimensionality Reduction
  - Converted Medical_keyword column into one.
  - Determined Correlation Coefficient for Continuous Column.
  - Performed MI (mutual information).
- Normalization

Northeastern University

Machine Learning Approaches

Classification:

- Logistic regression
- Random forest

Regressors:

- Support Vector Machine(SVM)
- XGBoost

Logistic regression

Logistic regression, despite its name, is a classification model rather than regression model.

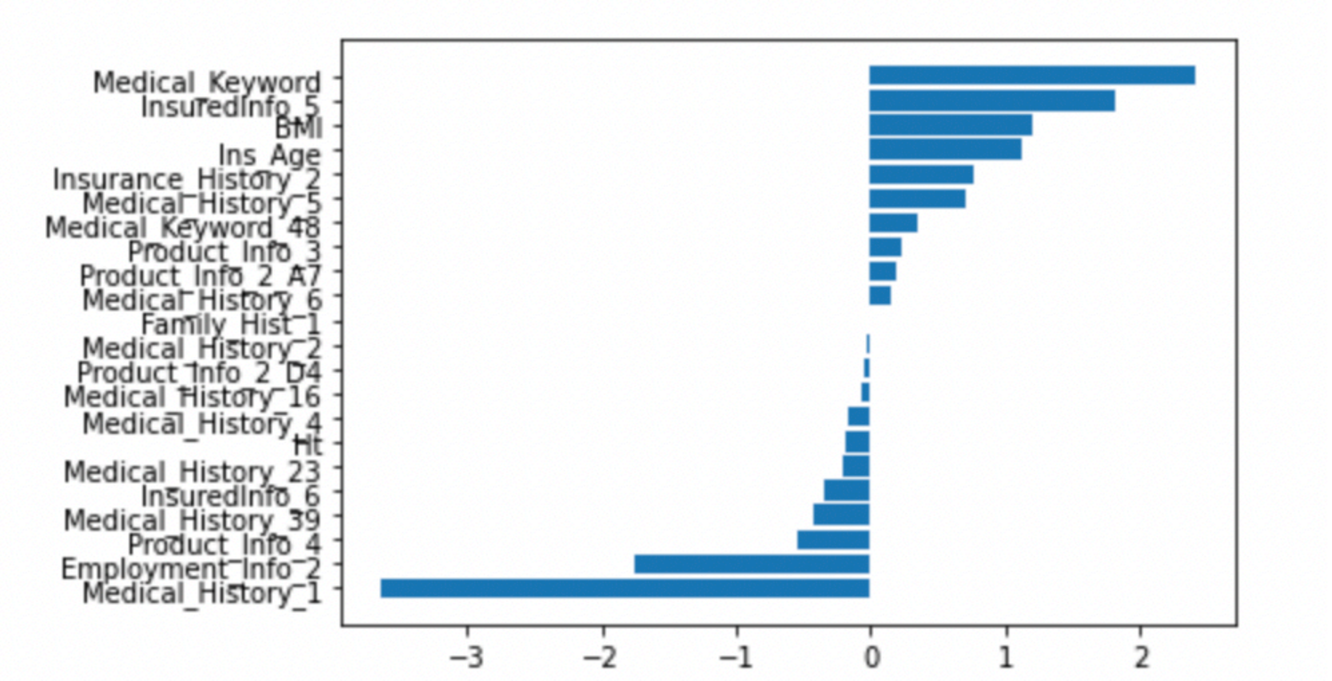
Logistic regression is a simple and more efficient method for binary and linear classification problems.

It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes.

Logistic regression

Methods Used: One-vs-rest (OvR)

Accuracy Score = 46%



Important Features

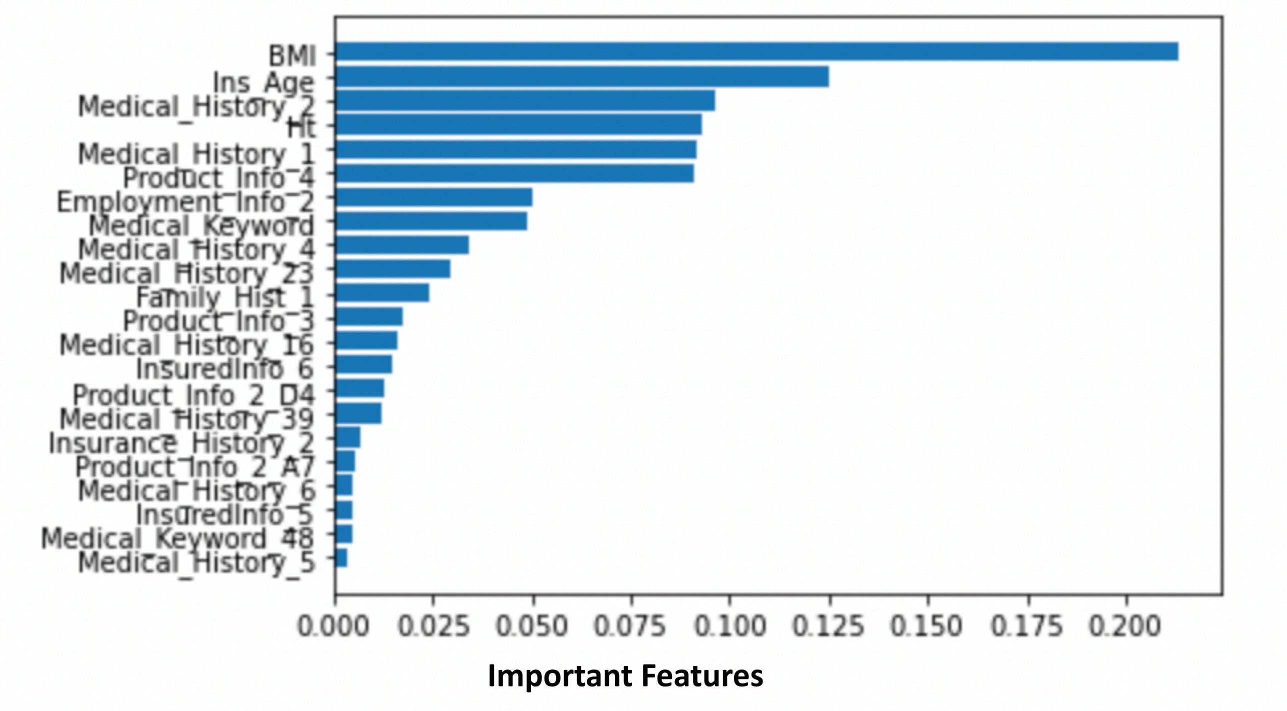
Random Forest

Random Forest is a tree-based machine learning algorithm that leverages the power of multiple decision trees for making decisions.

It does not rely on the feature importance given by a single decision tree.

Random Forest

Accuracy Score = 51.78%



Northeastern University

Support Vector Machine

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values.

Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value.

The threshold value is the distance between the hyperplane and boundary line.

Support Vector Machine

Methods Used

- Linear Kernel : The decision boundary is a straight line
  Accuracy Score = 32.3%
- Gaussian / RBF kernel : It projects the data into a Gaussian distribution

  c =1 & gamma = 0.1
  Accuracy Score = 33.6%

  c =1 & gamma = 1
  Accuracy Score = 32.9%

Northeastern University

XGBoost Regressor

XGBoost is a powerful approach for building supervised regression models.

The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners.

The objective function contains loss function and a regularization term.

XGBoost Regressor

- **RMSE:** It is the square root of mean squared error (MSE).
- **MAE:** It is an absolute sum of actual and predicted differences, but it lacks mathematically, that’s why it is rarely used, as compared to other metrics.

Accuracy Score = 32.81%

Mean Square error = 4.08

Northeastern University

Northeastern University

Conclusion and Findings

BMI is one of the most important feature for determining risk.

Random forest Classifier gave the highest accuracy.

Algorithm	Accuracy
Logistic Regression	46%
Random Forest Classifier	51.78%
Support Vector Machine (SVM)	33.6%
XGBoost Regressor	32.8%

Northeastern University

Thank You !