PREDICTING HEALTH INSURANCE PRICE FOR AN INDIVIDUAL OR FAMILY

Balaji H. Nalawade

INTRODUCTION

The majority of the countries finalize health insurance costs based on many factors such as age, number of people in families, etc.

What should be the actual health insurance price for an individual or a family is an issue for many companies?

We have already received samples required to perform all data analysis and machine learning tasks.

Now we have to perform all data analysis steps and finally create a machine learning model which can predict the health insurance cost.

HEALTH INSURANCE

OBJECTIVES

PROPOSAL IMPORTANCE

This proposal is very important in today's uncertain times. As a lot more advancement has happened in medical field, life expectancy has also increased. With all that medical treatment prices are also gone high. In this case our model can play big role.

GAP IN THE KNOWLEDGE

Data provided to us, it is very limited. There are many factors which can play vital role in this proposal. Medical conditions like diabetes or blood pressure, any prior surgeries, Medical checkup in last 3 or 6 months etc.

IMPORTANT FEATURES

To identify patterns in the data we have used countplot, distplot and histogram. By using multicollinearity by using VIF we got to know age and BMI columns are important features that may impact ML model.

ML MODELS

1. Linear Regression
2. XGBoost
3. Random Forest Regression
4. Support Vector Regression

STEPS INVOLVED IN IMPLEMENTATION

01 DATA ANALYSIS

02 ENCODE ALL CATEGORICAL DATA

03 DEALING WITH MISSING VALUES

04 EXAMINE MULTICOLLINEARITY

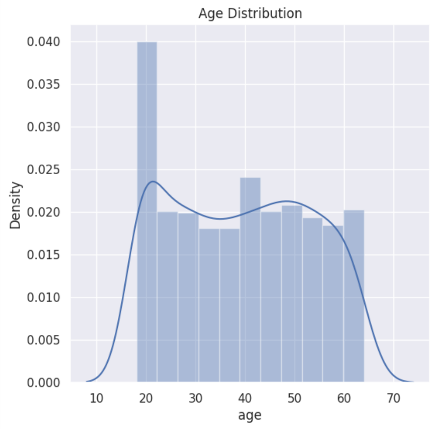
05 FEATURE SELECTION USING SELECTK BEST METHOD

06 LINEAR REGRESSION MODEL AND COST FUNCTION

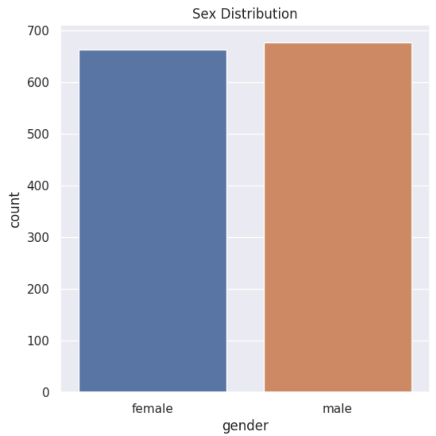
07 XGBOOST MODEL AND COST FUNCTION

08 RANDOM FOREST REGRESSION AND COST FUNCTION

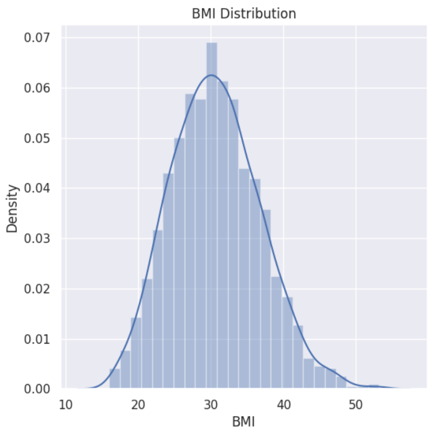
09 SUPPORT VECTOR REGRESSION AND COST FUNCTION



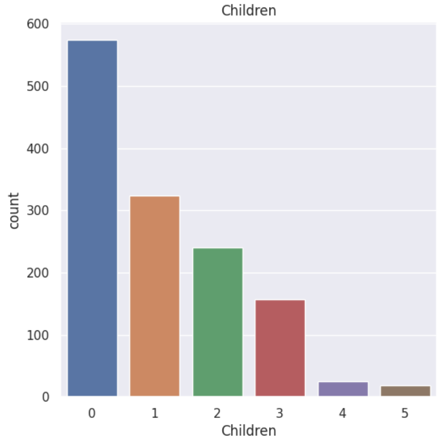
AGE COLUMN



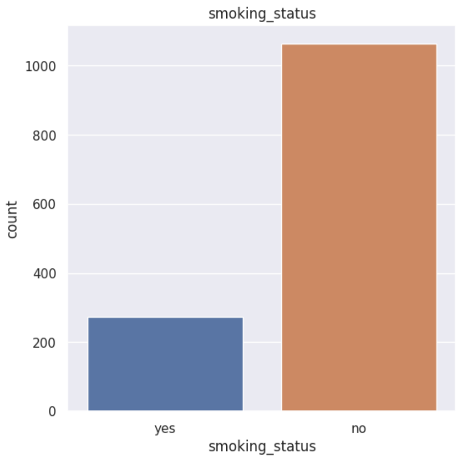
GENDER COLUMN



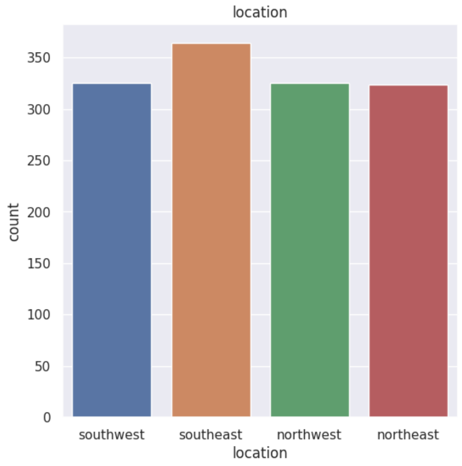
BMI DISTRIBUTION



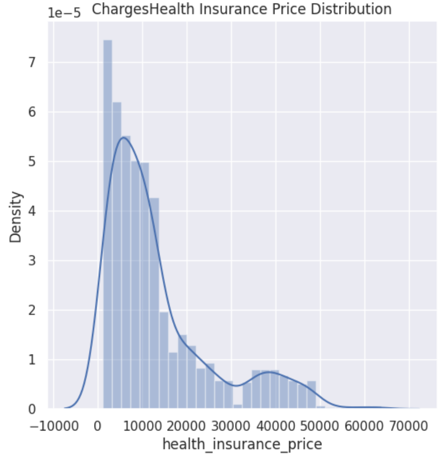
CHILDREN COLUMN



SMOKING STATUS COLUMN

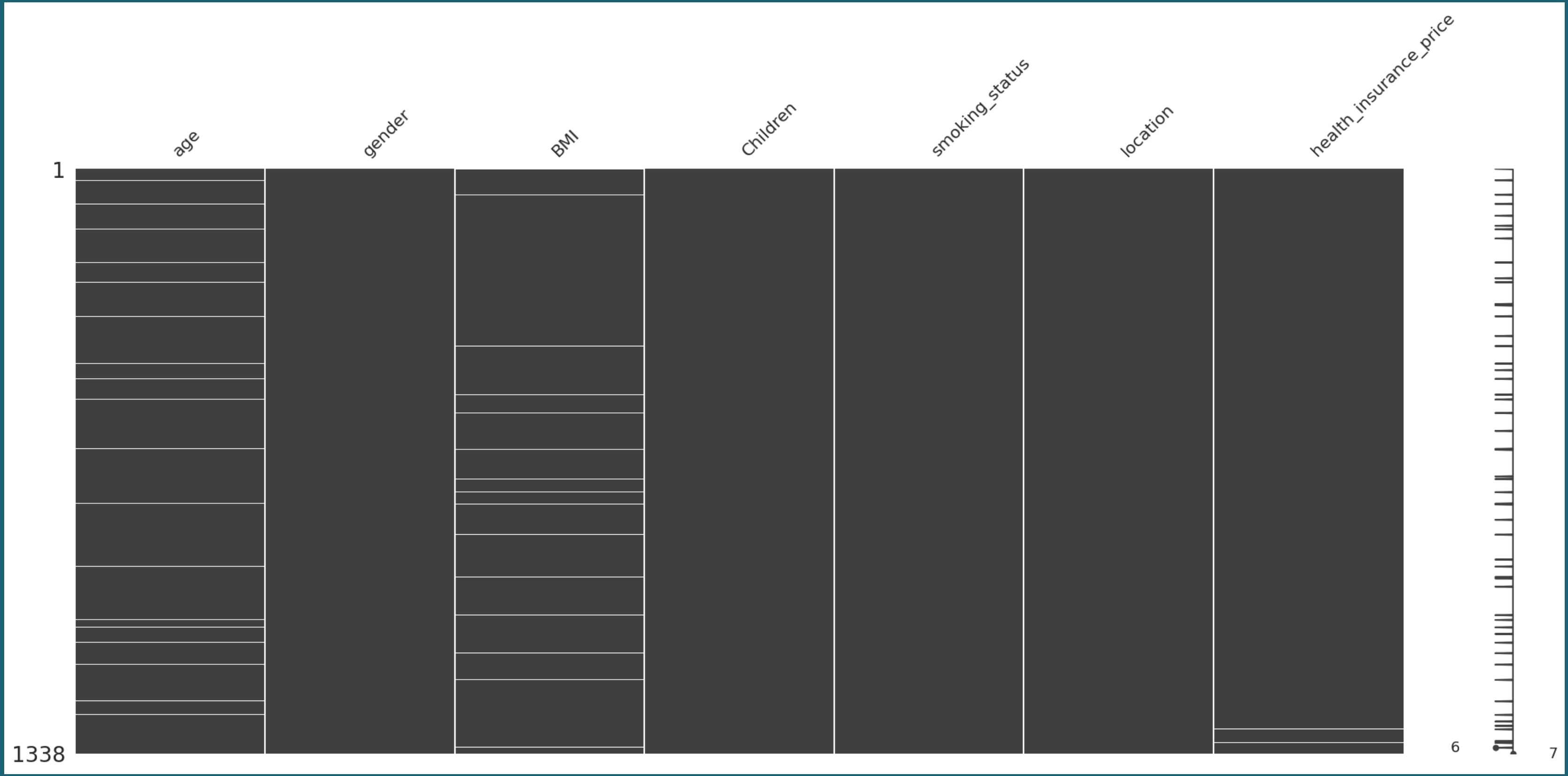


LOCATION COLUMN



DISTRIBUTION OF HEALTH_INSURANCE_PRICE

VISUALIZATION OF MISSING VALUES



ENCODE ALL CATEGORICAL DATA

```
# Ordinal encoding for location
# we can also use regular expression too
from sklearn.preprocessing import OrdinalEncoder
Or_enc = OrdinalEncoder()
insurance_dataset[["location"]] =
Or_enc.fit_transform(insurance_dataset[["location"]])
```

```
# label encoding for rest categorical variable
from sklearn.preprocessing import LabelEncoder

for col in ['gender','smoking_status']:
insurance_dataset[col] =
LabelEncoder().fit_transform(insurance_dataset[col])
```

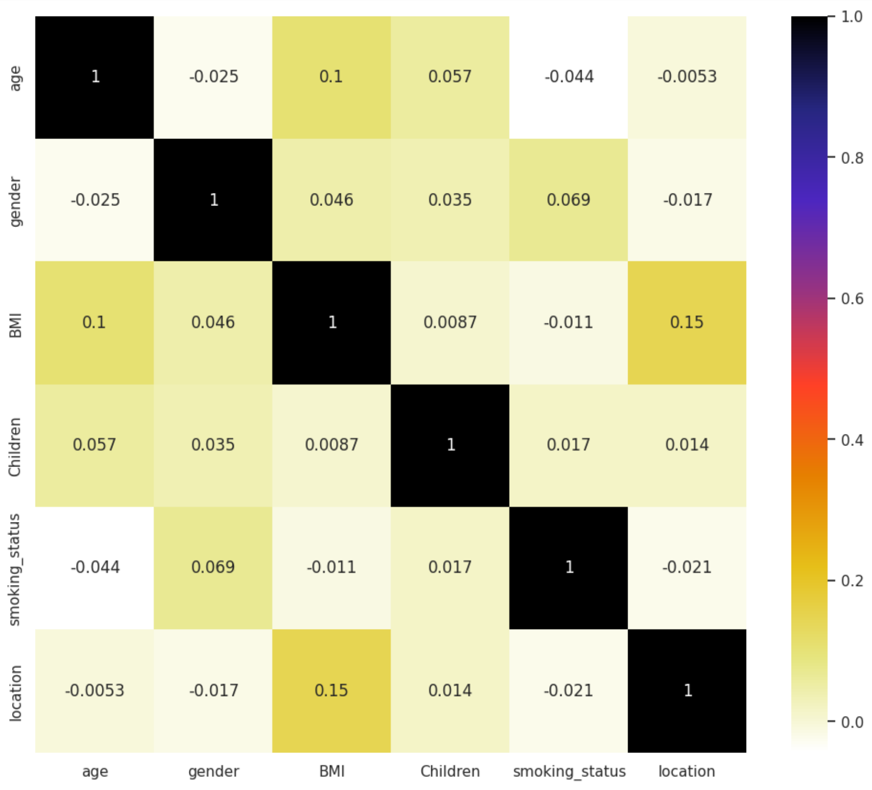
DEALING WITH MISSING VALUES

```
# Imputation using KNN
from fancyimpute import KNN
knn_imputer = KNN()
Independent_knn = Independent.copy(deep=True)
Independent_knn.iloc[:, :] =
knn_imputer.fit_transform(Independent_knn
```

```
# Imputation using MICE
from fancyimpute import IterativeImputer
MICE_imputer = IterativeImputer()
Independent_MICE = Independent.copy(deep=True)
Independent_MICE.iloc[:, :] =
MICE_imputer.fit_transform(Independent_MICE)
```

- FROM OVERALL EXPLORATION IT SEEMS THAT MICE AND KNN BOTH PERFORMED WELL
- HENCE, I WILL GO AHEAD WITH KNN IMPUTATION

PEARSON'S CORRELATION



THERE IS NO STRONG CORRELATION BETWEEN ANY TWO INDEPENDENT VARIABLE.

MULTICOLLINEARITY

```
# Examine multicollinearity using VIF
from statsmodels.stats.outliers_influence import
variance_inflation_factor
# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X_train.columns
# calculating VIF for each feature
vif_data["VIF"] =
[variance_inflation_factor(X_train.values, i)
for i in range(len(X_train.columns))]
print(vif_data)
```

- FROM MULTICOLLINEARITY WE GOT THAT AGE AND BMI COLUMN HAS HIGH VIF
- SO WE NEED TO DROP THESE TWO COLUMNS

COST FUNCTION VALUES

LINEAR REGRESSION

MAE:
5624.793157488833
MSE:
51171836.691327445
RMSE:
7153.449286276337.

XGBOOST

RMSE: 11257.635874

RANDOM FOREST

MAE: 5926.1186339416
MSE:
62429476.99969559
RMSE:
7901.232625337365

SUPPORT VECTOR REGRESSION

MAE:
8240.194990189058
MSE:
164568492.633928
95 RMSE:
12828.425181366922

CONCLUSION

Here, we performed 4 different models to check which model seems to give a better accuracy or least error. Overall, age and BMI do not seem to be a good predictor of a house price. Hence, they were dropped from all models. Linear regression seems to be the best model as it has the lowest error.

HEALTH INSURANCE

CONCLUSION

Here, we performed 4 different models to check which model seems to give a better accuracy or least error. Overall, age and BMI do not seem to be a good predictor of a house price. Hence, they were dropped from all models. Linear regression seems to be the best model as it has the lowest error.