

Assignment4 - Create Regression Model with Real Dataset & Model Deployment

1. Choose a real-world dataset that contains at least three features and a numeric target variable. that you want to predict. Ensure the dataset is suitable for regression analysis.

MIRI CHOI · UPDATED 6 YEARS AGO

2388

New Notebook

Download (16 kB)

Medical Cost Personal Datasets

Insurance Forecast by using Linear Regression

Data Card

Code (1247)

Discussion (12)

About Dataset

Context

Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

Content

Columns

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ²) using the ratio of height to weight, ideally 18.5 to 24.9

Usability 8.82

License
Database: Open Database, Cont...

Expected update frequency
Not specified



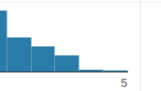
Tags
Education Health
Finance Insurance
Healthcare

insurance.csv (55.63 kB)

7 of 7 columns

About this file

This dataset consists of 1338 rows.

# age	sex	# bmi	# children	✓ smoker	region
Edad del asegurado	Género	Índice de masa corporal	Número de hijos	Indicador si fuma	Región de asegurado
	male 51% female 49%			0 0% 1 0% 2 0% 3 0% 4 0% 5 0%	southeas southwes Other (64
19	female	27.9	0	yes	southwe
18	male	33.77	1	no	southea
28	male	33	3	no	southea
33	male	22.705	0	no	northwe
32	male	28.88	0	no	northwe
31	female	25.74	0	no	southea
46	female	33.44	1	no	southea

Data Explorer

Version 1 (55.63 kB)

insurance.csv

Dataset: <https://www.kaggle.com/datasets/mirichoi0218/insurance>

2. Create a regression model to predict a target variable based on features from the dataset.

Step 1: Create a Linear Regression Model

```
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np

# Load the dataset
df = pd.read_csv("insurance.csv")

# Define the features and target
features = ['age', 'bmi', 'children']
target = 'charges'

X = df[features]
y = df[target]

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create and train the Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Calculate Mean Squared Error (MSE)
mse = mean_squared_error(y_test, y_pred)

# Calculate Root Mean Squared Error (RMSE)
rmse = np.sqrt(mse)

# Calculate R-squared (R^2)
r_squared = r2_score(y_test, y_pred)

print("Mean Squared Error (MSE):", mse)
print("Root Mean Squared Error (RMSE):", rmse)
print("R-squared (R^2):", r_squared)

Mean Squared Error (MSE): 131201335.64669806
Root Mean Squared Error (RMSE): 11454.315153980095
R-squared (R^2): 0.15489592484270753
```

Step 2: Save the Model

```
import joblib

# Save the model to a file
joblib.dump(model, 'linear_regression_model.pkl')
```

Step 3: Load the Model

```
import joblib

# Load the model
loaded_model = joblib.load('linear_regression_model.pkl')
```

Step 4: Show an Example of Prediction

```
import pandas as pd

# Example of making a prediction using the loaded model
example_data = pd.DataFrame({'age': [30], 'bmi': [25.0], 'children': [2]})
predicted_charges = loaded_model.predict(example_data)
print("Predicted Charges:", predicted_charges[0])
```

Predicted Charges: 9935.787545096708

Step 5: Create a Streamlit App with new file 'insurance_prediction_app.py'

```
import streamlit as st
import joblib
import pandas as pd
from sklearn.metrics import mean_squared_error
import numpy as np

# Load the saved model
loaded_model = joblib.load('linear_regression_model.pkl')

# Define the features for input
features = ['age', 'bmi', 'children']

# Define the Streamlit app
st.title("Insurance Charges Prediction")

# Create a sidebar for user input
st.sidebar.header('Enter Example Data')

age = st.sidebar.slider("Age", min_value=18, max_value=64, value=30)
bmi = st.sidebar.slider("BMI", min_value=15, max_value=50, value=25)
children = st.sidebar.slider("Number of Children", min_value=0, max_value=5, value=2)

# Create a DataFrame with the user input
example_data = pd.DataFrame([[age, bmi, children]], columns=features)

# Make predictions
predicted_charges = loaded_model.predict(example_data)

# Display the prediction
st.sidebar.header('Prediction')
st.sidebar.write(f"Predicted Charges: ${predicted_charges[0]:.2f}")

# Calculate Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)
actual_charges = 10000 # Replace with the actual charges if available

if actual_charges:
    mse = mean_squared_error([actual_charges], [predicted_charges[0]])
    rmse = np.sqrt(mse)
    st.sidebar.header('Model Evaluation Metrics')
    st.sidebar.write(f"MSE: {mse:.2f}")
    st.sidebar.write(f"RMSE: {rmse:.2f}")
else:
    st.sidebar.write("Actual charges not provided. Unable to calculate MSE and RMSE.")

# Main content area
st.header('Insurance Charges Prediction')
```

3. Deploy the regression model as a web application using Streamlit.io.

[← Back](#)

Deploy an app

Repository

[Paste GitHub URL](#)

taohoo3698/predictChanges

Branch

main

Main file path

insurance_prediction_app.py

App URL (Optional)

predictchanges-63130500015

.streamlit.app

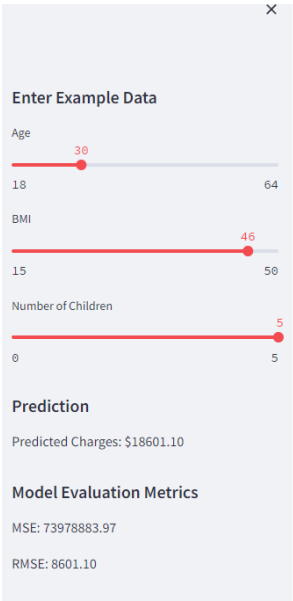
Domain is available

[Advanced settings...](#)

Deploy!

4. Put the evaluation result on your web app.

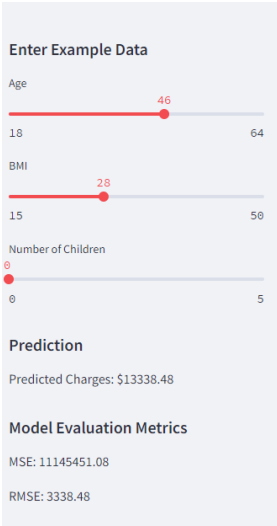
Result 1



Insurance Charges Prediction

Insurance Charges Prediction

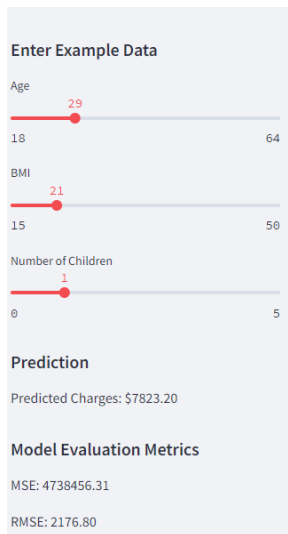
Result 2



Insurance Charges Prediction

Insurance Charges Prediction

Result 3



Enter Example Data

Age: 29 (range 18 to 64)

BMI: 21 (range 15 to 50)

Number of Children: 1 (range 0 to 5)

Prediction

Predicted Charges: \$7823.20

Model Evaluation Metrics

MSE: 4738456.31

RMSE: 2176.80

Insurance Charges Prediction

Insurance Charges Prediction

5. In the Streamlit Sharing settings for your app, make sure to set it to "Public" or "Anyone with the link can view."

Link: <https://predictchanges-63130500015.streamlit.app/>