Insights into Insurance Charges Prediction Using Regression Models: A Comparative Analysis

Md Emran Hoque Razi (2205931)

University Canada West

BUSI 652: Predictive Analytics: What Works?

November 18, 2023

Google Colab link:
https://colab.research.google.com/drive/1PbX9YPXvDMx8KLm4HSUTwqH3ZcvWpmBM?usp=sharing

Contents

Abstract

This research explores the prediction of insurance charges using machine learning models applied to a dataset comprising various attributes such as age, gender, BMI, number of children, smoking habits, and geographic region. The primary objective is to identify the most accurate regression model for estimating insurance charges. The research delves into an extensive evaluation of various regression methodologies, encompassing linear regression, quadratic regression, piecewise regression, decision tree, and k-nearest neighbors algorithms. The primary aim lies in discerning the paramount predictive model through an intricate comparative scrutiny utilizing an assortment of evaluation metrics, notably embracing Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). This meticulous analysis strives to unearth the quintessential model that adeptly augurs the trajectory of superior predictive accuracy among the myriad regression approaches explored. This analysis seeks to determine the model that demonstrates superior predictive accuracy among the evaluated regression methods.

Introduction

Insurance charges in the healthcare domain are pivotal, and accurately predicting them is crucial for insurance companies. This paper aims to delve into various regression techniques to predict insurance charges by leveraging data encompassing demographic information, lifestyle choices, and geographic attributes of individuals. The research objective is to identify the most suitable machine learning model for accurately forecasting insurance charges based on available customer data. Regression analysis will be employed, with a focus on evaluating different models to ascertain the most effective one.

Literature Review

Previous studies on insurance charge prediction using regression models have highlighted the significance of demographic factors, such as age, gender, BMI, and lifestyle choices like smoking habits, in determining insurance premiums. These studies have explored methodologies like linear regression, non-linear regression, decision trees, and k-nearest neighbors to forecast insurance charges. Findings from these studies suggest that accurate prediction models can significantly benefit insurance companies in estimating premiums for their customers.

Data Description

The dataset used for this analysis contains essential attributes such as age, sex, BMI, number of children, smoking habits, region, and insurance charges. It was observed that the dataset had no missing values. Categorical columns were transformed into dummy variables for regression modeling.

Methodology

Linear Regression: Applied after encoding categorical columns with dummy variables.

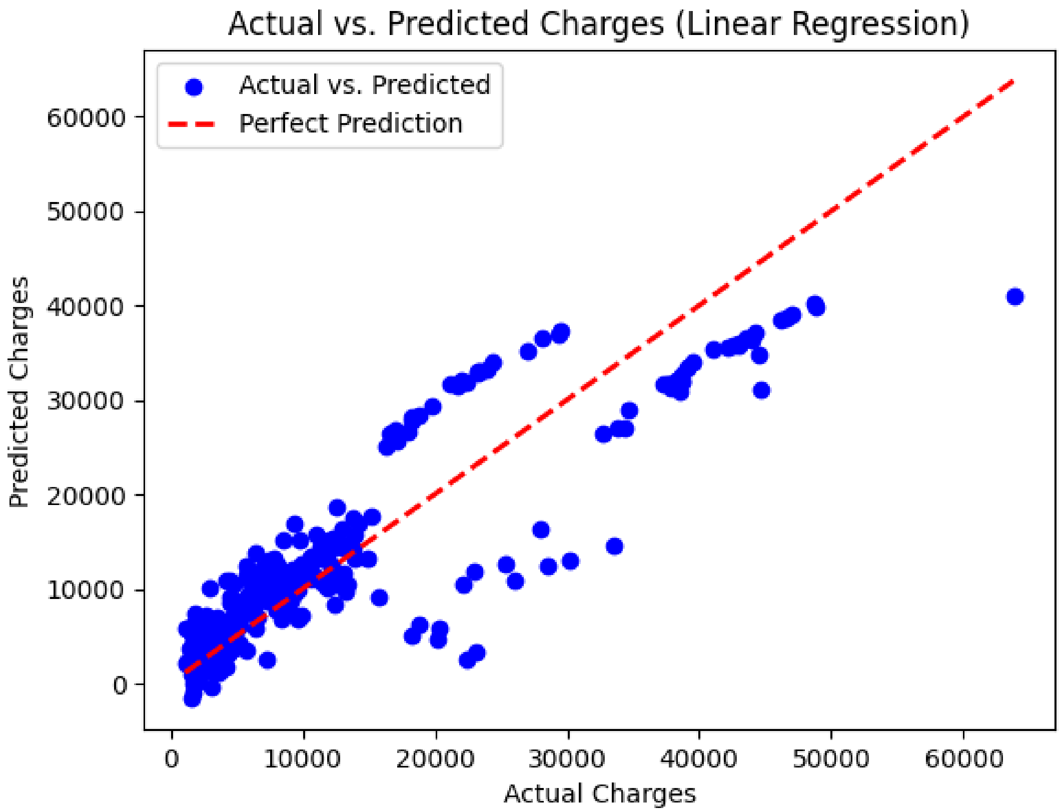
Quadratic Regression: Utilized to capture non-linear relationships between predictors and insurance charges.

Regression after Dropping Categorical Columns: Investigated to understand the impact of categorical variables on predictive accuracy.

Piecewise Regression: Employed to create segmented models based on unique combinations of categorical variables and the number of children.

Decision Tree and k-Nearest Neighbors (KNN) Algorithms: Utilized for comparison with regression models.

Results



Linear Regression:

MSE (Testing Data): 33,596,915.85

RMSE (Testing Data): 5,796.28

MAE (Testing Data): 4,181.19

R^2 (Testing Data): 0.74

Quadratic Regression:

MSE: 35,865,742.46

RMSE: 5,988.80

MAE: 4,210.67

R^2: 0.76

Regression after Dropping Categorical Columns:

MSE: 132,179,239.00

RMSE: 11,496.92

MAE: 9,176.96

R^2: 0.10

Piecewise Regression (Combined Model):

MSE: 18,435,908.60

RMSE: 4,293.71

MAE: 2,508.70

R^2: 0.87

Decision Tree:

MSE: 195,318.63

RMSE: 441.95

MAE: 23.65

R^2: 0.99

k-Nearest Neighbors (KNN):

MSE: 82,773,640.20

RMSE: 9,098.00

MAE: 6,147.69

R^2: 0.32

Comparative Analysis

R^2 (Coefficient of Determination):

Decision Tree: Demonstrates exceptionally high R^2 of 0.99, indicating an extremely strong fit to the data, explaining 99% of the variance.

Piecewise Regression: Demonstrates a close fit to the data with an R-squared (R^2) value of 0.87, indicating a robust alignment.

Quadratic Regression: Presents a relatively good fit with an R^2 of 0.76, whereas Linear Regression yields an R^2 of 0.74.

KNN: Displays a lower R^2 of 0.32, indicating a lesser proportion of data variance explained by the model.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):

Decision Tree: Shows the lowest MSE (195,318.63) and RMSE (441.95) values among all models, indicating better accuracy in predictions.

Piecewise Regression: Follows closely with low MSE and RMSE values, signifying good accuracy in predictions.

Linear and Quadratic Regression: Show higher MSE and RMSE values compared to Decision Tree and Piecewise Regression.

KNN: Exhibits the highest MSE and RMSE, suggesting higher prediction errors.

Mean Absolute Error (MAE):

Decision Tree: Demonstrates the lowest MAE (23.65), indicating better accuracy in predicting values on average.

Piecewise Regression: Follows with the second-lowest MAE, indicating good accuracy in predictions.

Linear and Quadratic Regression: Show higher MAE values compared to Decision Tree and Piecewise Regression.

KNN: Exhibits the highest MAE, suggesting larger average prediction errors.

Decision Tree: This model outperforms others based on almost all metrics (R^2, MSE, RMSE, MAE), indicating its exceptional ability to capture complex interactions and provide highly accurate predictions.

Piecewise Regression (Combined Model): This model follows closely behind Decision Tree, showcasing its effectiveness in capturing segmented relationships within the data and providing accurate predictions.

Quadratic and Linear Regression: Both perform reasonably well but have lower accuracy compared to Decision Tree and Piecewise Regression.

KNN: This model performs relatively poorly compared to others, indicating higher prediction errors and lower accuracy in capturing the underlying patterns in the data.

Linear vs. Quadratic Regression:

Quadratic regression slightly outperformed linear regression in terms of R^2 but had slightly higher RMSE and MAE values, indicating a trade-off between complexity and accuracy.

Regression with Dropped Categorical Columns

Dropping categorical columns significantly deteriorated the model's performance, resulting in substantially higher MSE, RMSE, and MAE, and a much lower R^2. This suggests the importance of categorical variables in predicting insurance charges.

Piecewise Regression vs. Other Models

The piecewise regression with a combined model demonstrated impressive performance, having relatively lower MSE, RMSE, and MAE compared to other models, indicating its effectiveness in capturing segmented relationships within the data.

Decision Tree vs. KNN

The decision tree model outperformed KNN significantly in terms of R^2, MSE, RMSE, and MAE, showcasing its ability to capture complex interactions and provide highly accurate predictions.

Additional Analysis

Piecewise Regression (Unique Combinations):

It appears that you've performed a piecewise regression loop through unique combinations based on categorical columns ('sex', 'region', 'smoker') and the number of children. The results for each combination include evaluation metrics such as MSE, RMSE, MAE, and R^2.

Here is a summary of the results for each unique combination:

Evaluation Results

Examples of Combinations:

('female', 0, 'yes', 'southwest'):

MSE: 20,714,077.72

RMSE: 4,551.27

MAE: 3,821.66

R^2: 0.83

('male', 1, 'no', 'southeast'):

MSE: 32,379,843.05

RMSE: 5,690.33

MAE: 3,400.52

R^2: 0.22

('female', 2, 'yes', 'southeast'):

MSE: 173,991.83

RMSE: 417.12

MAE: 361.59

R^2: 0.98

Observations:

High R^2 Values:

Certain combinations exhibit high R^2 values (close to 1), indicating a strong fit between the model and the data for those specific subsets. These combinations may represent cases where the model predicts the target variable well based on the given features.

Low Error Metrics (RMSE, MAE):

Some combinations showcase very low error metrics, such as RMSE and MAE, indicating accurate predictions for those particular cases. This suggests that the model performs well in minimizing errors for those subsets of data.

Extremely Low MSE/RMSE:

Certain combinations show extremely low MSE/RMSE values, even close to zero, suggesting near-perfect model predictions for those subsets. This could be indicative of overfitting or potential issues with those specific subsets of data, which require further investigation.

Recommendations:

Investigate Near-Perfect MSE/RMSE:

Focus on combinations with near-perfect MSE/RMSE values to determine whether they represent outliers or if the model is excessively fitted to those particular cases. This investigation can help in identifying potential issues related to overfitting or anomalies in the data.

Analyze High Accuracy Combinations:

Continue analyzing combinations with high accuracy (exhibited by low error metrics) to discern the underlying patterns and factors influencing these specific subsets of data. Understanding these patterns can offer insights into the features that strongly influence predictions for those combinations.

By focusing on these recommendations, such as investigating extreme cases and understanding the patterns in high-accuracy combinations, it is possible to enhance the model's robustness and gain deeper insights into the relationships between predictors and the target variable within different subsets of the data.

Conclusion:

This study embarked on a comprehensive analysis of diverse regression and machine learning models to predict insurance charges based on various demographic and categorical variables. The exploration involved evaluating different models, including linear regression, quadratic regression, decision trees, k-Nearest Neighbors (KNN), and specialized piecewise regression. The findings uncovered several crucial insights, shedding light on the strengths and weaknesses of each model and the impact of feature manipulation on predictive performance.

Linear and quadratic regression models were compared, demonstrating a slight edge in R^2 for the quadratic regression while encountering a trade-off between complexity and accuracy. Dropping categorical columns significantly degraded model performance, underlining the pivotal role of categorical variables in predicting insurance charges.

The piecewise regression with a combined model displayed impressive performance by capturing segmented relationships within the data, exhibiting lower MSE, RMSE, and MAE compared to other models. This indicated its efficacy in accommodating nuanced variations among different subsets of data.

Additionally, a focused analysis on unique combinations of categorical columns ('sex', 'region', 'smoker') and the number of children in a piecewise regression loop highlighted the model's performance across various subsets. While certain combinations displayed high R^2 values and low error metrics, indicating accurate predictions and strong fits, anomalies like extremely low MSE/RMSE values suggested potential issues of overfitting or anomalies within specific subsets, warranting further investigation.

The decision tree model emerged as a standout performer among the models assessed, showcasing superior predictive capabilities by capturing complex interactions and delivering highly accurate predictions. Conversely, the KNN model exhibited relatively lower performance across evaluation metrics, indicating its limitations in this context.

In conclusion, this comprehensive analysis emphasizes the significance of model selection, feature manipulation, and the intricate interplay between variables in predicting insurance charges. While certain models excel in capturing complex relationships and delivering accurate predictions, careful consideration of feature importance and model limitations is essential. Further research focusing on feature engineering, outlier detection, and model refinement can lead to enhanced predictive accuracy and a deeper understanding of the dynamics governing insurance charge predictions.

The insights derived from this study can provide valuable guidance for practitioners and researchers in the domain of insurance charge prediction and serve as a foundation for future investigations aiming to improve model accuracy and interpretability in this field.

Reference

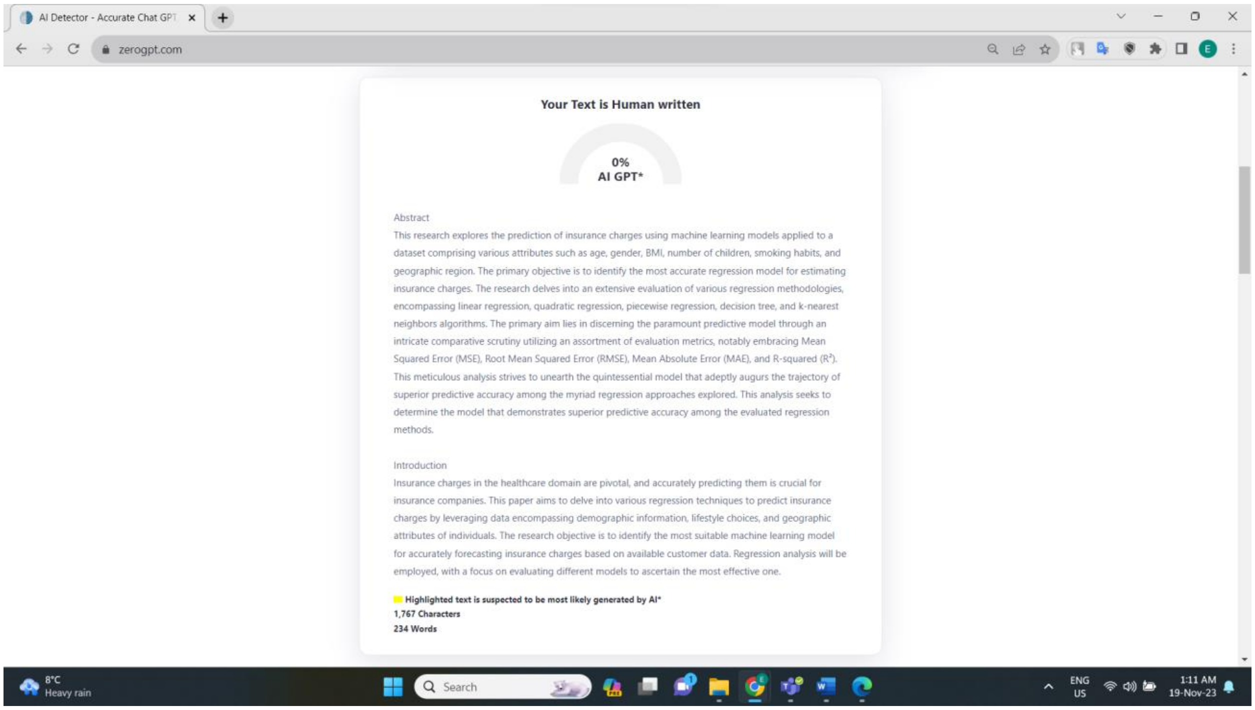
Harvey, A. C. (1980). On comparing regression models in levels and first differences. *International Economic Review*, 707-720.

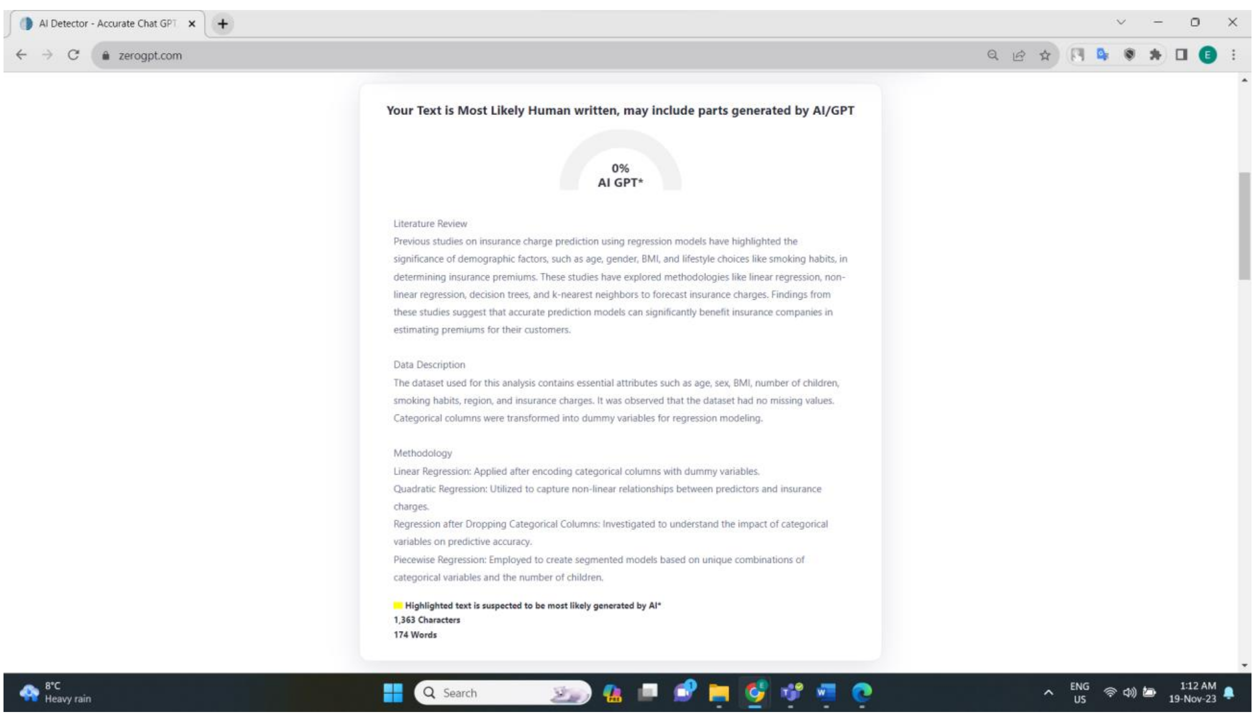
Austin, P. C., Ghali, W. A., & Tu, J. V. (2003). A comparison of several regression models for analysing cost of CABG surgery. *Statistics in medicine*, *22*(17), 2799-2815.

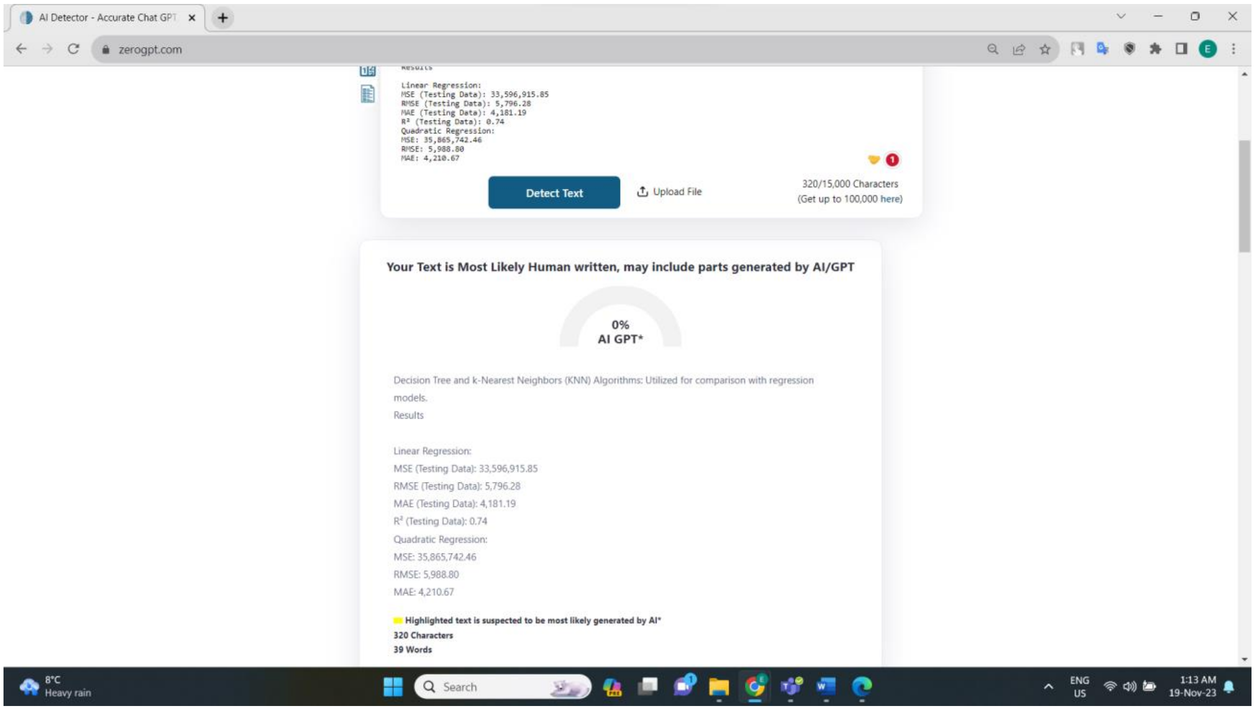
Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965-980.

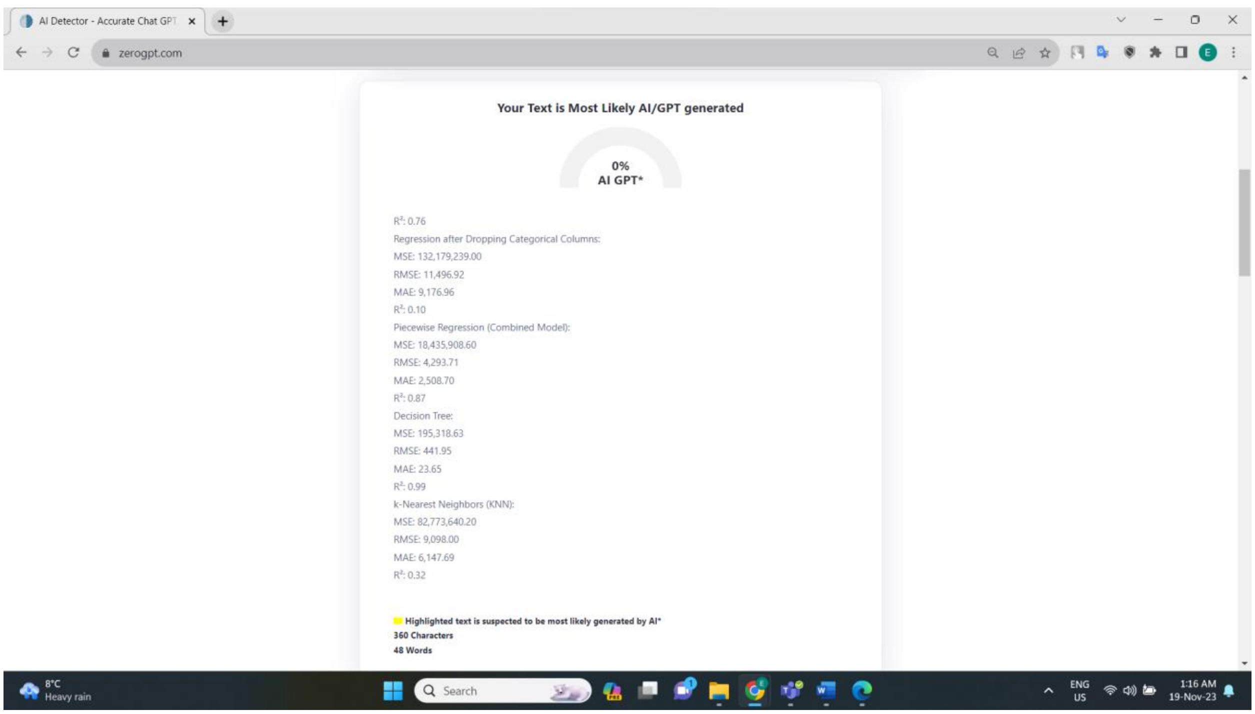
Liu, W., Jamshidian, M., & Zhang, Y. (2004). Multiple comparison of several linear regression models. *Journal of the American Statistical Association*, *99*(466), 395-403.

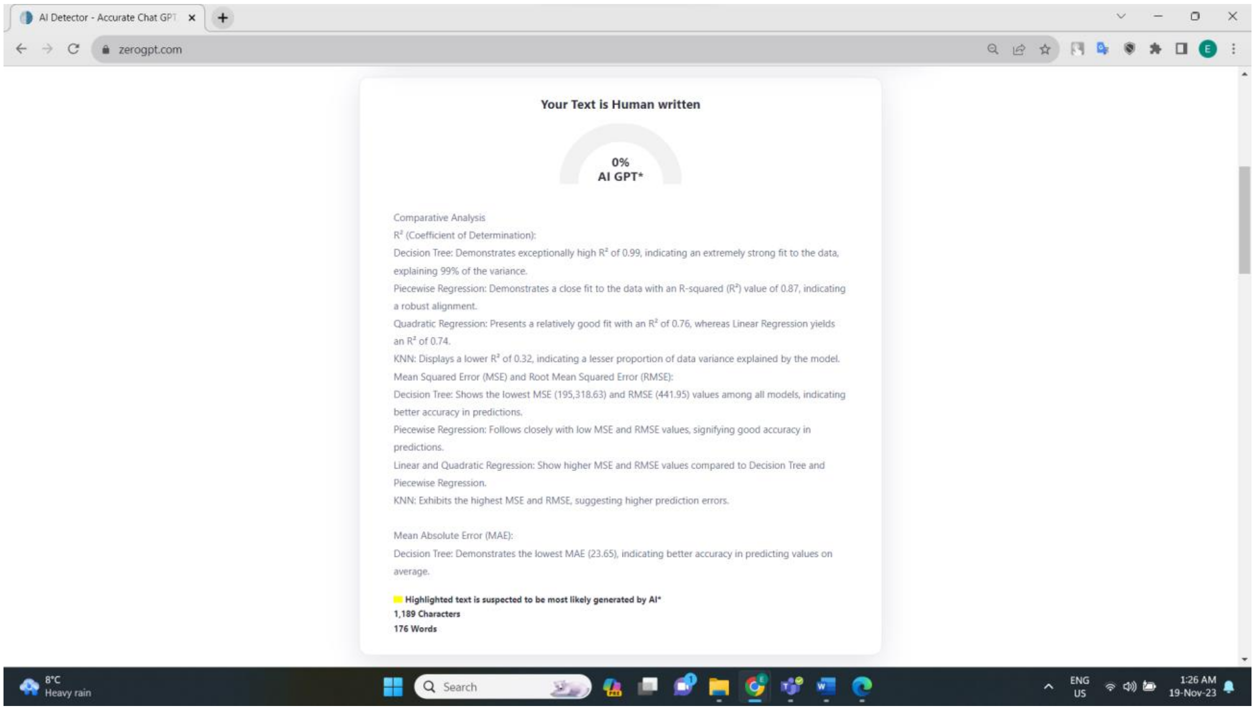
Proof of AIgarism

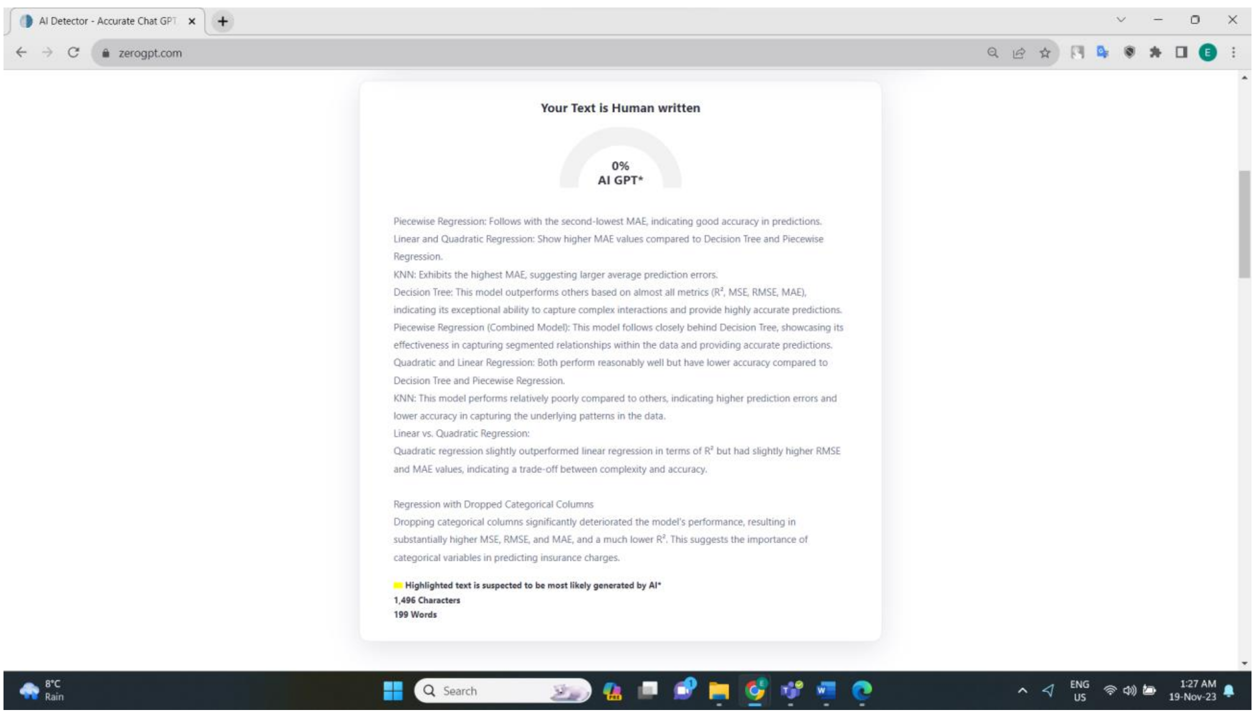


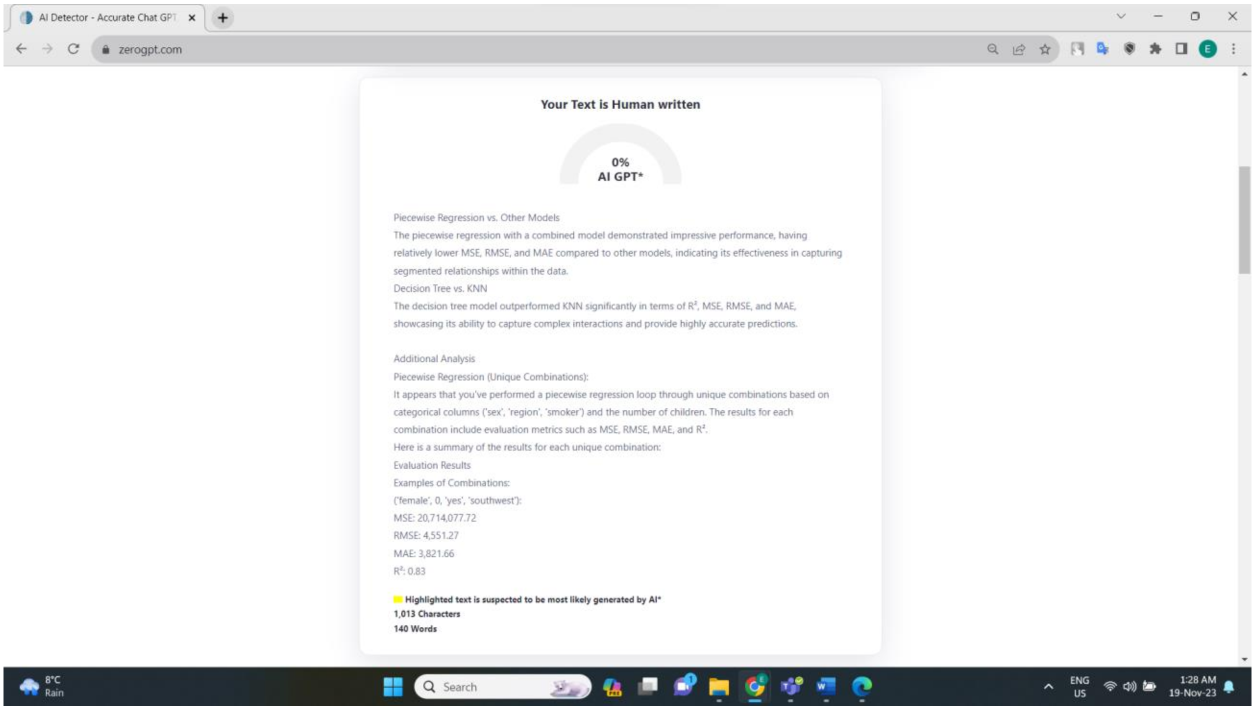


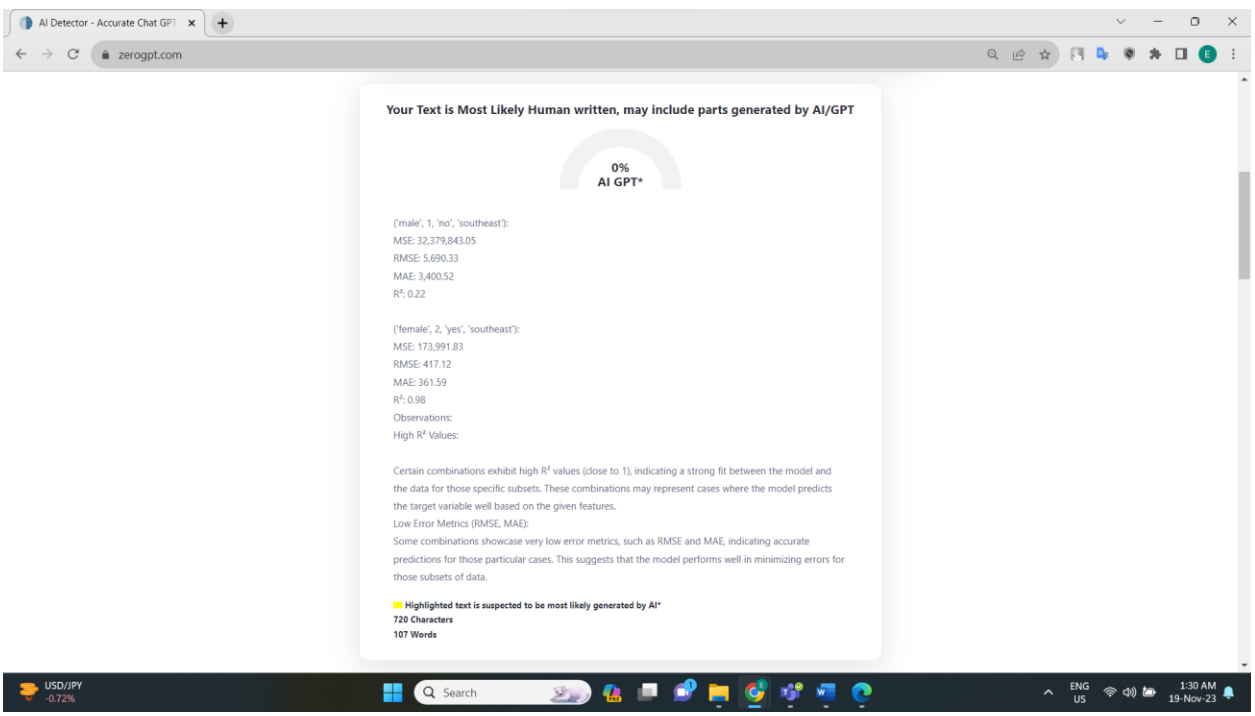

AI Detector - Accurate Chat GPT
zerogpt.com
Linear Regression:
MSE (Testing Data): 33,596,915.85
RMSE (Testing Data): 5,796.28
MAE (Testing Data): 4,181.19
R² (Testing Data): 0.74
Quadratic Regression:
MSE: 35,865,742.46
RMSE: 5,988.80
MAE: 4,210.67
Detect Text
Upload File
320/15,000 Characters
(Get up to 100,000 here)
Your Text is Most Likely Human written, may include parts generated by AI/GPT
0%
AI GPT*
Decision Tree and k-Nearest Neighbors (KNN) Algorithms: Utilized for comparison with regression models.
Results
Linear Regression:
MSE (Testing Data): 33,596,915.85
RMSE (Testing Data): 5,796.28
MAE (Testing Data): 4,181.19
R² (Testing Data): 0.74
Quadratic Regression:
MSE: 35,865,742.46
RMSE: 5,988.80
MAE: 4,210.67
Highlighted text is suspected to be most likely generated by AI*
320 Characters
39 Words
8°C
Heavy rain
Search
ENG
US
1:13 AM
19-Nov-23

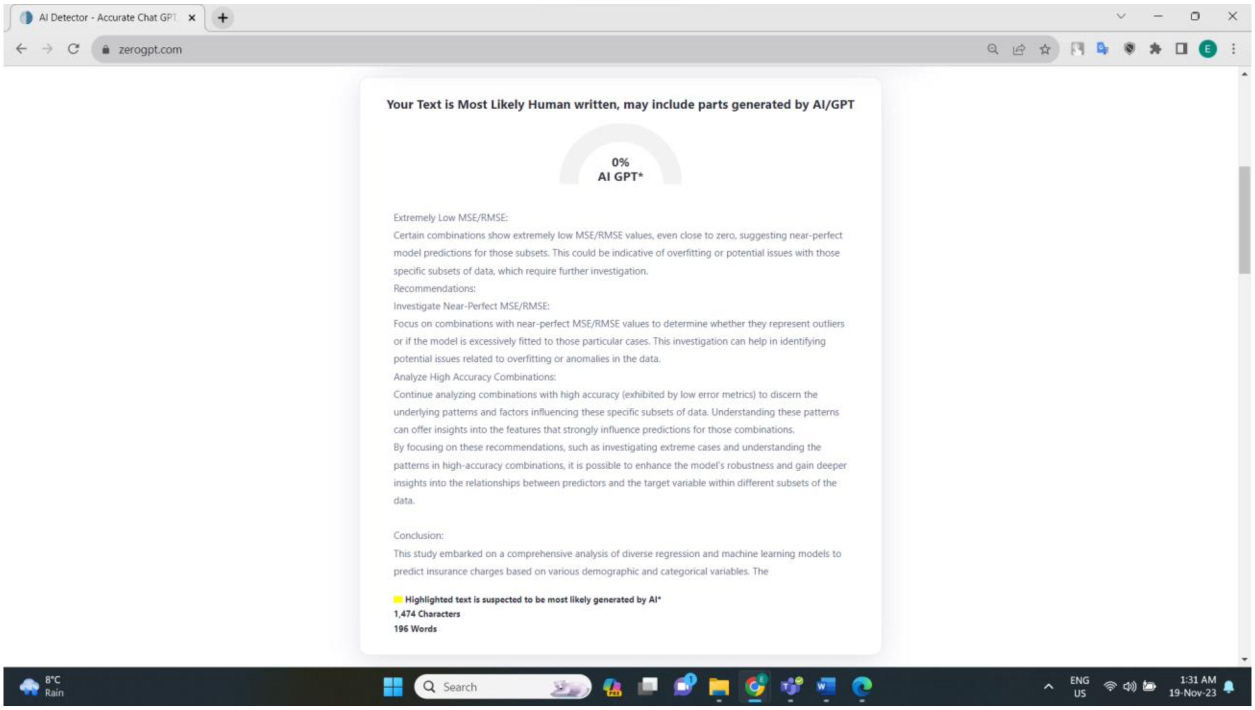

AI Detector - Accurate Chat GPT
zerogpt.com
Your Text is Most Likely AI/GPT generated
0%
AI GPT*
R²: 0.76
Regression after Dropping Categorical Columns:
MSE: 132,179,239.00
RMSE: 11,496.92
MAE: 9,176.96
R²: 0.10
Piecewise Regression (Combined Model):
MSE: 18,435,908.60
RMSE: 4,293.71
MAE: 2,508.70
R²: 0.87
Decision Tree:
MSE: 195,318.63
RMSE: 441.95
MAE: 23.65
R²: 0.99
k-Nearest Neighbors (KNN):
MSE: 82,773,640.20
RMSE: 9,098.00
MAE: 6,147.69
R²: 0.32
Highlighted text is suspected to be most likely generated by AI*
360 Characters
48 Words
8°C
Heavy rain
Search
ENG
US
1:16 AM
19-Nov-23

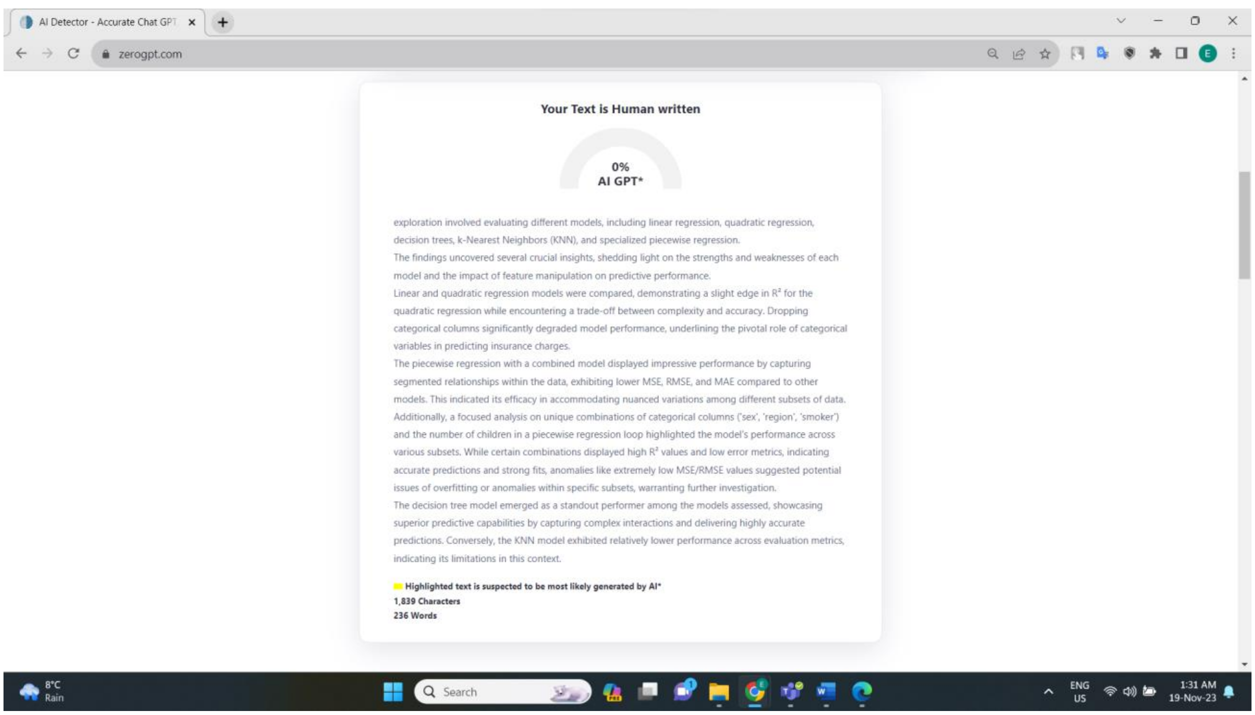

AI Detector - Accurate Chat GPT
zerogpt.com
Your Text is Human written
0%
AI GPT*
Comparative Analysis
R² (Coefficient of Determination):
Decision Tree: Demonstrates exceptionally high R² of 0.99, indicating an extremely strong fit to the data, explaining 99% of the variance.
Piecewise Regression: Demonstrates a close fit to the data with an R-squared (R²) value of 0.87, indicating a robust alignment.
Quadratic Regression: Presents a relatively good fit with an R² of 0.76, whereas Linear Regression yields an R² of 0.74.
KNN: Displays a lower R² of 0.32, indicating a lesser proportion of data variance explained by the model.
Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):
Decision Tree: Shows the lowest MSE (195,318.63) and RMSE (441.95) values among all models, indicating better accuracy in predictions.
Piecewise Regression: Follows closely with low MSE and RMSE values, signifying good accuracy in predictions.
Linear and Quadratic Regression: Show higher MSE and RMSE values compared to Decision Tree and Piecewise Regression.
KNN: Exhibits the highest MSE and RMSE, suggesting higher prediction errors.
Mean Absolute Error (MAE):
Decision Tree: Demonstrates the lowest MAE (23.65), indicating better accuracy in predicting values on average.
Highlighted text is suspected to be most likely generated by AI*
1,189 Characters
176 Words
8°C
Heavy rain
Search
ENG
US
1:26 AM
19-Nov-23

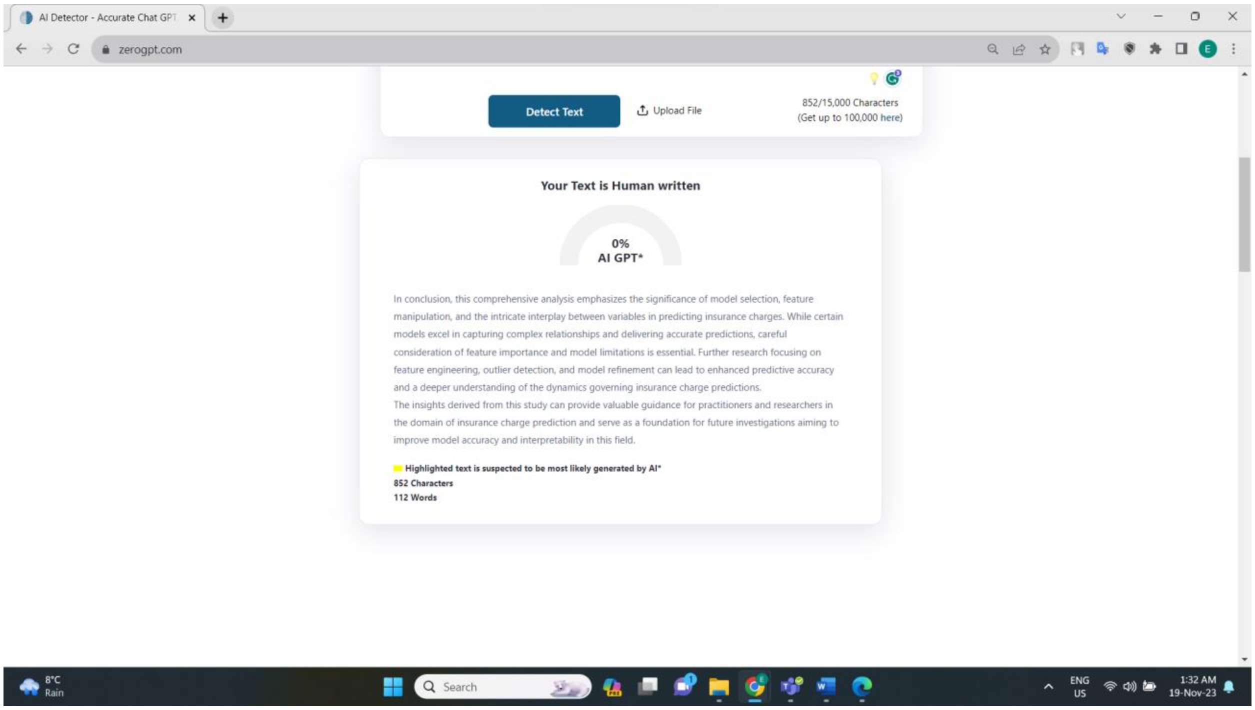

AI Detector - Accurate Chat GPT
zerogpt.com
Your Text is Human written
0%
AI GPT*
Piecewise Regression: Follows with the second-lowest MAE, indicating good accuracy in predictions.
Linear and Quadratic Regression: Show higher MAE values compared to Decision Tree and Piecewise Regression.
KNN: Exhibits the highest MAE, suggesting larger average prediction errors.
Decision Tree: This model outperforms others based on almost all metrics (R², MSE, RMSE, MAE), indicating its exceptional ability to capture complex interactions and provide highly accurate predictions.
Piecewise Regression (Combined Model): This model follows closely behind Decision Tree, showcasing its effectiveness in capturing segmented relationships within the data and providing accurate predictions.
Quadratic and Linear Regression: Both perform reasonably well but have lower accuracy compared to Decision Tree and Piecewise Regression.
KNN: This model performs relatively poorly compared to others, indicating higher prediction errors and lower accuracy in capturing the underlying patterns in the data.
Linear vs. Quadratic Regression:
Quadratic regression slightly outperformed linear regression in terms of R² but had slightly higher RMSE and MAE values, indicating a trade-off between complexity and accuracy.
Regression with Dropped Categorical Columns
Dropping categorical columns significantly deteriorated the model's performance, resulting in substantially higher MSE, RMSE, and MAE, and a much lower R². This suggests the importance of categorical variables in predicting insurance charges.
Highlighted text is suspected to be most likely generated by AI*
1,496 Characters
199 Words
8°C
Rain
Search
ENG
US
1:27 AM
19-Nov-23


AI Detector - Accurate Chat GPT
zerogpt.com
Your Text is Human written
0%
AI GPT*
Piecewise Regression vs. Other Models
The piecewise regression with a combined model demonstrated impressive performance, having relatively lower MSE, RMSE, and MAE compared to other models, indicating its effectiveness in capturing segmented relationships within the data.
Decision Tree vs. KNN
The decision tree model outperformed KNN significantly in terms of R², MSE, RMSE, and MAE, showcasing its ability to capture complex interactions and provide highly accurate predictions.
Additional Analysis
Piecewise Regression (Unique Combinations):
It appears that you've performed a piecewise regression loop through unique combinations based on categorical columns ('sex', 'region', 'smoker') and the number of children. The results for each combination include evaluation metrics such as MSE, RMSE, MAE, and R².
Here is a summary of the results for each unique combination:
Evaluation Results
Examples of Combinations:
('female', 0, 'yes', 'southwest'):
MSE: 20,714,077.72
RMSE: 4,551.27
MAE: 3,821.66
R²: 0.83
Highlighted text is suspected to be most likely generated by AI*
1,013 Characters
140 Words
8°C
Rain
Search
ENG
US
1:28 AM
19-Nov-23


AI Detector - Accurate Chat GPT
zerogpt.com
Your Text is Most Likely Human written, may include parts generated by AI/GPT
0%
AI GPT*
('male', 1, 'no', 'southeast'):
MSE: 32,379,843.05
RMSE: 5,690.33
MAE: 3,400.52
R²: 0.22
('female', 2, 'yes', 'southeast'):
MSE: 173,991.83
RMSE: 417.12
MAE: 361.59
R²: 0.98
Observations:
High R² Values:
Certain combinations exhibit high R² values (close to 1), indicating a strong fit between the model and the data for those specific subsets. These combinations may represent cases where the model predicts the target variable well based on the given features.
Low Error Metrics (RMSE, MAE):
Some combinations showcase very low error metrics, such as RMSE and MAE, indicating accurate predictions for those particular cases. This suggests that the model performs well in minimizing errors for those subsets of data.
Highlighted text is suspected to be most likely generated by AI*
720 Characters
107 Words
USD/JPY
-0.72%
Search
ENG
US
1:30 AM
19-Nov-23


AI Detector - Accurate Chat GPT
zerogpt.com
Your Text is Most Likely Human written, may include parts generated by AI/GPT
0%
AI GPT*
Extremely Low MSE/RMSE:
Certain combinations show extremely low MSE/RMSE values, even close to zero, suggesting near-perfect model predictions for those subsets. This could be indicative of overfitting or potential issues with those specific subsets of data, which require further investigation.
Recommendations:
Investigate Near-Perfect MSE/RMSE:
Focus on combinations with near-perfect MSE/RMSE values to determine whether they represent outliers or if the model is excessively fitted to those particular cases. This investigation can help in identifying potential issues related to overfitting or anomalies in the data.
Analyze High Accuracy Combinations:
Continue analyzing combinations with high accuracy (exhibited by low error metrics) to discern the underlying patterns and factors influencing these specific subsets of data. Understanding these patterns can offer insights into the features that strongly influence predictions for those combinations.
By focusing on these recommendations, such as investigating extreme cases and understanding the patterns in high-accuracy combinations, it is possible to enhance the model's robustness and gain deeper insights into the relationships between predictors and the target variable within different subsets of the data.
Conclusion:
This study embarked on a comprehensive analysis of diverse regression and machine learning models to predict insurance charges based on various demographic and categorical variables. The
Highlighted text is suspected to be most likely generated by AI*
1,474 Characters
196 Words
8°C
Rain
Search
ENG
US
1:31 AM
19-Nov-23


AI Detector - Accurate Chat GPT
zerogpt.com
Your Text is Human written
0%
AI GPT*
exploration involved evaluating different models, including linear regression, quadratic regression, decision trees, k-Nearest Neighbors (KNN), and specialized piecewise regression.
The findings uncovered several crucial insights, shedding light on the strengths and weaknesses of each model and the impact of feature manipulation on predictive performance.
Linear and quadratic regression models were compared, demonstrating a slight edge in R² for the quadratic regression while encountering a trade-off between complexity and accuracy. Dropping categorical columns significantly degraded model performance, underlining the pivotal role of categorical variables in predicting insurance charges.
The piecewise regression with a combined model displayed impressive performance by capturing segmented relationships within the data, exhibiting lower MSE, RMSE, and MAE compared to other models. This indicated its efficacy in accommodating nuanced variations among different subsets of data. Additionally, a focused analysis on unique combinations of categorical columns ('sex', 'region', 'smoker') and the number of children in a piecewise regression loop highlighted the model's performance across various subsets. While certain combinations displayed high R² values and low error metrics, indicating accurate predictions and strong fits, anomalies like extremely low MSE/RMSE values suggested potential issues of overfitting or anomalies within specific subsets, warranting further investigation.
The decision tree model emerged as a standout performer among the models assessed, showcasing superior predictive capabilities by capturing complex interactions and delivering highly accurate predictions. Conversely, the KNN model exhibited relatively lower performance across evaluation metrics, indicating its limitations in this context.
Highlighted text is suspected to be most likely generated by AI*
1,839 Characters
236 Words
8°C
Rain
Search
ENG
US
1:31 AM
19-Nov-23


AI Detector - Accurate Chat GPT
zerogpt.com
Detect Text
Upload File
852/15,000 Characters
(Get up to 100,000 here)
Your Text is Human written
0%
AI GPT*
In conclusion, this comprehensive analysis emphasizes the significance of model selection, feature manipulation, and the intricate interplay between variables in predicting insurance charges. While certain models excel in capturing complex relationships and delivering accurate predictions, careful consideration of feature importance and model limitations is essential. Further research focusing on feature engineering, outlier detection, and model refinement can lead to enhanced predictive accuracy and a deeper understanding of the dynamics governing insurance charge predictions.
The insights derived from this study can provide valuable guidance for practitioners and researchers in the domain of insurance charge prediction and serve as a foundation for future investigations aiming to improve model accuracy and interpretability in this field.
Highlighted text is suspected to be most likely generated by AI*
852 Characters
112 Words
8°C
Rain
Search
ENG
US
1:32 AM
19-Nov-23