Stanford Summer Session 2024

DATASCI-112: Principles of Data Science

Project Name: Smart N'Sure

Team name: Peanuts

Discussion section: 06

Team members:

Bruna Alves Maziero | bmaziero@stanford.edu

Gunjan Siddharth | gunjan03@stanford.edu

Ho Shing Louis Lau | louislau@stanford.edu

Introduction

Insurance pricing in real life is largely influenced by various risk factors associated with the policyholder. Insurers use these factors to assess the likelihood and potential cost of a claim, and set premiums accordingly. Common variables include age, health status, and lifestyle choices.

Our project sought to delve deeper into the dynamics of health insurance pricing by analyzing relationships among these variables using machine learning techniques. We employed models to predict, cluster, and categorize the variables, enhancing our understanding of the insurance pricing mechanism.

For our analysis, we utilized a public domain dataset from Kaggle titled "Insurance Dataset for Predicting Health Insurance Premiums in the US." This dataset comprises 1,000,000 rows and includes 12 key variables:

Numerical:

- Age: The age of the insured individual.
- Charges: The health insurance charges for the individual.
- BMI (Body Mass Index): A measure of body fat based on height and weight.
- Children: The number of children covered by the insurance plan.

Categorical:

- Gender: The gender of the insured individual.
- Smoking Status: Indicates whether the individual is a smoker.
- Region: The geographical region of the insured individual.
- Medical History: Information about the individual's old medical problems.
- Family Medical History: Information about the family's medical record.
- Exercise Frequency: The frequency of the individual's exercise routine.
- Occupation: The occupation of the insured individual.
- Coverage Level: The type of insurance plan.

The dataset was exceptionally clean; the only modification required was addressing missing values in 'medical_history' and 'family_medical_history' columns. These were categorized under "No record" to denote an absence of medical history.

This refined data preparation allowed us to proceed with an accurate analysis, ensuring the reliability of our findings.

To achieve our goal, we proposed four analytical questions, which are designed to enhance our understanding and analysis of our main inquiry: What factors most significantly influence health care insurance pricing? The questions we posed were:

I. Which features are the most critical in determining insurance pricing?
II. What is the predicted price of insurance given the different characteristics of a patient? How do Linear Regression and K Nearest Neighbors compare in terms of predicting insurance charges, and which model performs the best?
III. Classify the insurance type (Premium, Standard, Basic) based on different characteristics without knowing the charges.
IV. Can we identify distinct clusters of insurance prices based on policyholders' health and demographic characteristics? What are the defining features of each cluster?

The questions are meant to be answered in this sequence because it allows for a logical progression from identifying the most influential variables to applying and comparing predictive models. This approach not only facilitates a deeper understanding of the factors driving insurance costs but also tests the predictive power of our models in various scenarios, culminating in a comprehensive segmentation of the insurance market based on tangible data insights.

Given the large size of the dataset, processing all 1,000,001 rows would have been computationally intensive and time-consuming. To manage this more efficiently, we decided to use a 50% random sample of the data. While this approach reduced the computational load, it still allowed us to capture significant patterns and trends relevant to our analysis, ensuring the insights gained were both valuable and reliable.

Methodology

To effectively explore the dynamics of health insurance pricing, our team devised a methodology centered around four key analytical questions. Each question is designed to progressively build upon the insights garnered from its predecessor, facilitating a comprehensive understanding of what drives insurance costs. Our project adopted a systematic approach, addressing these primary analytical questions step by step, with each phase building on the findings of the previous one to create a layered understanding of the factors influencing insurance costs.

As a first step in addressing the questions, we transformed the features by one-hot encoding the categorical features and standardizing the numeric features.

Analytical Questions:

1. Identifying Critical Features Our first objective was to determine which features are the most critical in influencing insurance pricing. This involved an analysis using a univariate regression model to understand the top features that influence the insurance charge. We then evaluated the Linear Regression model on every combination of the 10 features and found the optimal set of features. This step was crucial for setting the groundwork for more complex analyses in subsequent stages.

2. Predicting Insurance Pricing Next, we delved into predicting the price of insurance based on various patient characteristics, comparing the performance of Linear Regression and K Nearest Neighbors (KNN) models in predicting insurance charges. This comparison aimed to identify which model offered the most reliable predictions, incorporating all dataset features and employing a training-validation-testing split of 70-20-10. The models were evaluated using metrics like RMSE, MSE, and SSE, with a focus on RMSE for its effectiveness in highlighting prediction errors.

3. Classification of Insurance Types Our third question focused on classifying the type of insurance (Premium, Standard, Basic) based solely on the characteristics of

the policyholder, excluding the charges. For this classification task, we utilized all available features and developed a KNN classification model. The model was integrated into a pipeline that included preprocessing steps and model training. The evaluation metrics used were accuracy, precision, recall, and the F1 score, supported by a confusion matrix to visually assess the classification outcomes.

4. Clustering of Insurance Prices Finally, we explored the potential to identify distinct clusters of insurance prices based on the health and demographic characteristics of policyholders, excluding coverage level from the clustering variables. We applied a K-means clustering algorithm to the standardized and encoded dataset. The goal was to discover and define the features characterizing each cluster. The clustering effectiveness was assessed using the silhouette score, and we conducted a detailed analysis of cluster centroids and feature distributions to interpret the defining characteristics.

Analysis & Results

Our analysis began with an exploratory data analysis (EDA) to understand the relationships between various factors and insurance charges. Following the EDA, we addressed the first two analytical questions, which focused on identifying the critical features influencing insurance pricing and comparing predictive models.

I. Exploratory Data Analysis (EDA)

The EDA aimed to uncover how different variables such as age, gender, occupation, and medical history impact insurance charges. The key findings from this analysis were:

Smoking Status: Smoking was identified as a significant factor in determining insurance charges. Smokers generally face higher premiums, reflecting the increased health risks associated with smoking.

Medical History: Both personal and family medical histories, particularly involving heart disease, were found to significantly impact insurance costs. Individuals with a history of heart disease tend to incur higher charges, emphasizing the role of pre-existing conditions in insurance pricing.

Age and Occupation: Age also played a crucial role, with older individuals generally facing higher insurance costs. Occupation influenced charges, with white-collar workers typically paying more than those in other job categories.

These insights set the stage for more detailed analyses in the subsequent questions, providing a foundational understanding of the factors most closely associated with insurance pricing.

II. Question 1: Identifying Critical Features for Insurance Pricing

The first analytical question aimed to identify which features are most critical in determining insurance pricing. To achieve this, we conducted univariate analysis by fitting linear regression models to each feature individually and assessing their impact on insurance charges.

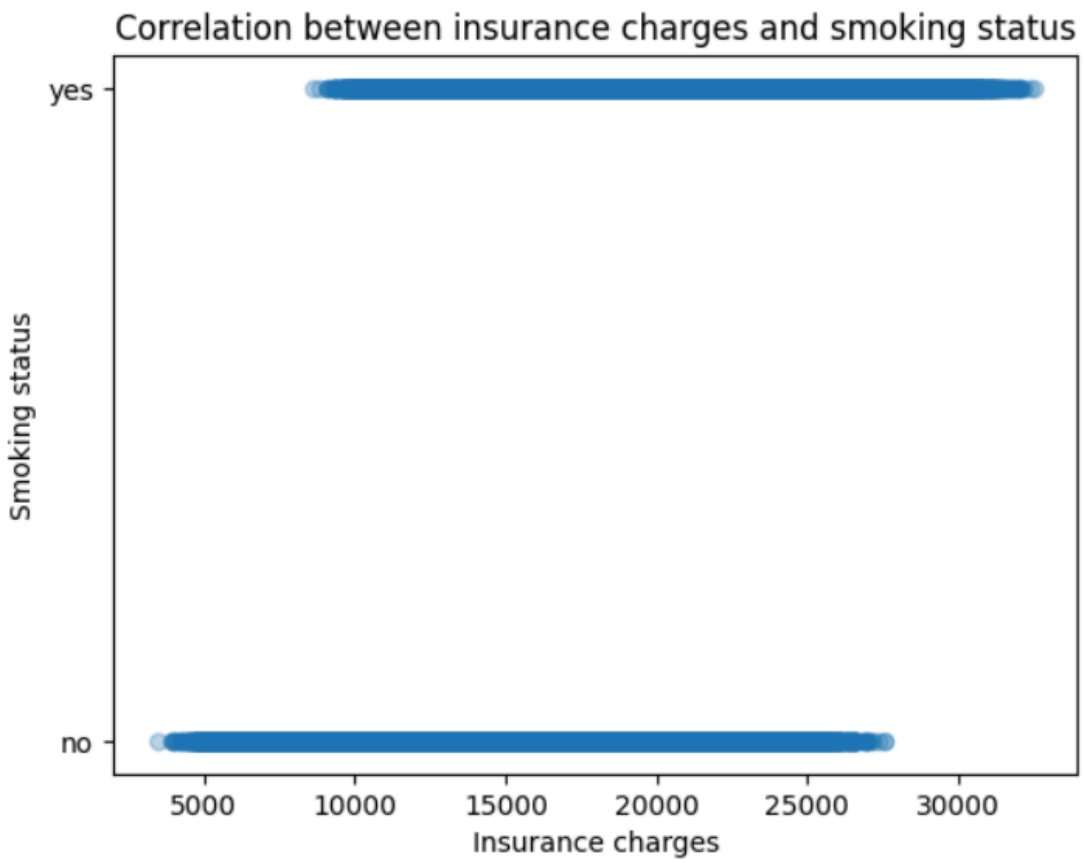

Fig. 1. Smoking Status vs. Insurance Charges

Key Features Identified:

Smoking Status: As seen in Fig. 1 above, this was the most critical feature, with a strong positive correlation with insurance charges. Smokers consistently paid higher premiums, making this the top predictor in our models.

Medical History: The presence of heart disease in either personal or family medical history was also a strong predictor of higher insurance costs. This underscores the significance of chronic health conditions in determining insurance pricing.

Occupation and Age: Occupation and age were also important, with older individuals and those in white-collar jobs generally facing higher charges.

The univariate analysis highlighted these factors as the most influential in determining insurance costs, providing a clear direction for model development in subsequent analyses.

Combinations of features: We also evaluated the Linear Regression model using all kinds of feature combinations using 10 of them. We found that the combination of all 10 features gave the most optimal prediction.

III. Question 2: Predicting Insurance Pricing Using Linear Regression and KNN

In the second analytical question, we aimed to predict insurance prices based on various patient characteristics, comparing the performance of Linear Regression and K Nearest Neighbors (KNN) models.

Model Comparison:

Linear Regression: This model performed well, leveraging the linear relationships between the features and insurance charges. It provided a straightforward, interpretable model with reasonably accurate predictions.

K Nearest Neighbors (KNN): While KNN offered a non-linear approach, it was slightly less effective than Linear Regression in predicting insurance charges. This was particularly evident in scenarios where the relationship between the variables and charges was more linear.

Evaluation Metrics:

The models were evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Sum of Squared Errors (SSE), with RMSE being the primary focus for its ability to highlight prediction errors.

Linear Regression slightly outperformed KNN, particularly in cases where the relationships between features and insurance charges were linear. The RMSE values

indicated that Linear Regression was more consistent in predicting insurance costs, making it the preferred model for this analysis.

In conclusion, the EDA and subsequent analyses in Question 1 and Question 2 revealed that smoking status, medical history, and occupation are the most critical factors influencing insurance pricing. Linear Regression emerged as the most effective model for predicting insurance charges based on these features.

IV. Question 3: Classification of Coverage Levels using KNN:

The KNN (k=5) model's performance is suboptimal with an accuracy of around 33.2%. This suggests that the model is not effectively distinguishing between the coverage levels. Eg: Precision: 0.33 - of all instances predicted as Premium, 33% are actually Premium. Recall: 0.46 - Of all actual Premium instances, 46% were correctly predicted. Recall measures how well the model identifies positive instances.

Precision, recall, and F1-Scores are all quite low across the classes, particularly for the Basic class, which indicates issues with identifying this class.

V. Question 4: Clustering the patients characteristics into different insurance levels

We applied KMeans with 3 clusters and randomized initiation. The **cluster purity** was extremely low at 6e-06. This indicates that the cluster contains a mix of different coverage_level categories.

We also graphed the clusters (as seen in Fig. 2 below) against age and BMI which gave three distinct class colors. While the cluster purity score is very low, the distinct coloring in the graph suggests that KMeans has successfully grouped the data based on the two features, but these groups do not align well with the coverage_level classes.

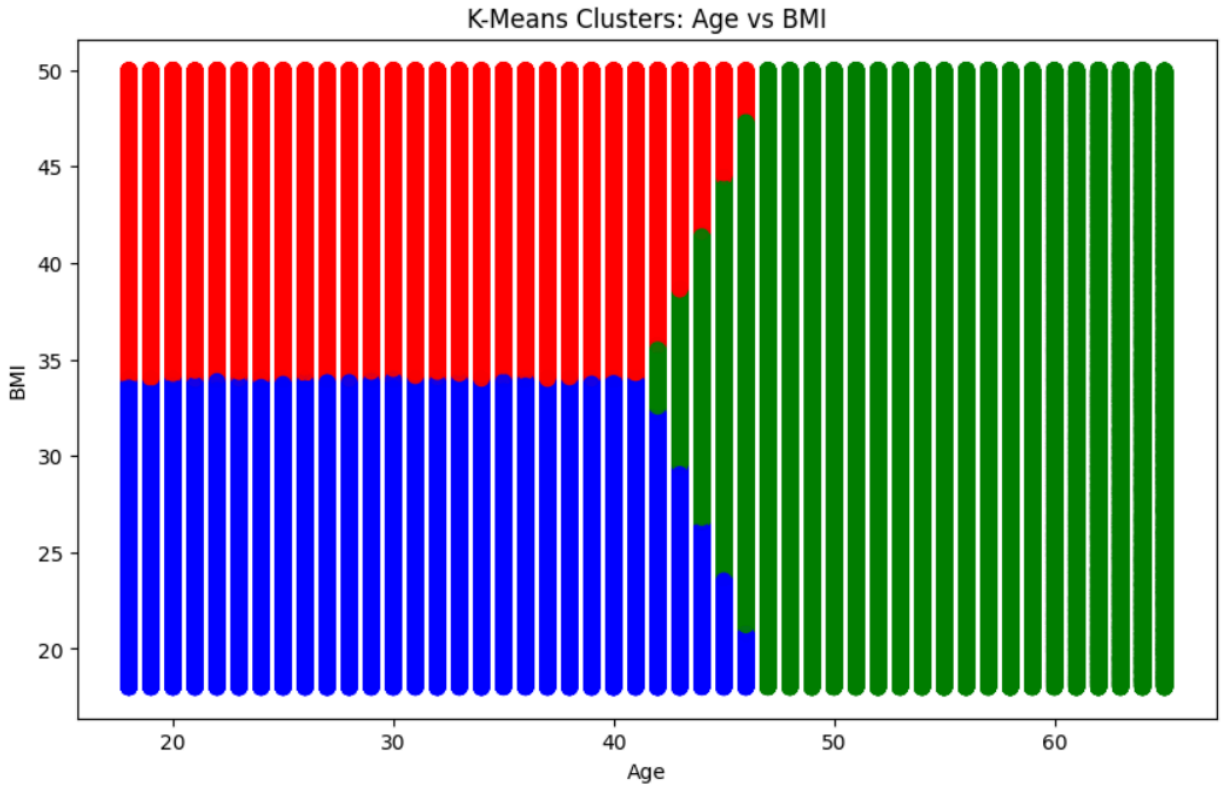

Fig. 2. KNN clusters

Discussion and Conclusions

This analysis aimed to uncover the factors influencing health insurance pricing, predict costs, classify coverage levels, and explore potential clusters within the data. While significant insights were gained, challenges were encountered, particularly in classification and clustering.

Critical Factors in Pricing:

Smoking status, medical history, occupation, and age emerged as the primary drivers of insurance costs. Linear Regression confirmed these findings, showing strong predictive power when using all selected features.

Prediction of Insurance Costs:

Linear Regression outperformed KNN in predicting insurance prices, suggesting that the relationships between features and costs are predominantly linear.

Coverage Level Classification:

KNN struggled with classifying coverage levels, achieving only 33.2% accuracy. This indicates a need for either more complex models or better-suited features to effectively differentiate between coverage levels.

Clustering Analysis:

KMeans clustering, while visually distinct when plotted against age and BMI, produced a very low cluster purity score. This suggests that the clusters formed do not align well with coverage levels, highlighting the complexity of the data.

Conclusion:

While the analysis successfully identified key factors in insurance pricing and demonstrated the effectiveness of Linear Regression for prediction, the classification and clustering tasks revealed areas for improvement. Future efforts should focus on enhancing classification models and exploring better clustering approaches to capture the nuances in the data.