PGP - Data Science and Business Analysis - Jan’ 21

Capstone Notes part 1

*by Sarang Manohar*



Table of Contents

[Table of Contents 2](#_Toc91449024)

[Introduction of the business problem 3](#_Toc91449025)

[a) Defining the problem statement 3](#_Toc91449026)

[b) Need of the study/project 3](#_Toc91449027)

[c) Understanding business/social opportunity 3](#_Toc91449028)

[Data Report 3](#_Toc91449029)

[a) Understanding how data was collected in terms of time, frequency and methodology 3](#_Toc91449030)

[b) Visual inspection of data (rows, columns, descriptive details) 3](#_Toc91449031)

[c) Understanding of attributes (variable info, renaming if required) 4](#_Toc91449032)

[Exploratory data analysis 5](#_Toc91449033)

[a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) 5](#_Toc91449034)

[b) Bivariate analysis (relationship between different variables, correlations) 10](#_Toc91449035)

[c) Removal of unwanted variables (if applicable) 19](#_Toc91449036)

[d) Missing Value treatment (if applicable) 19](#_Toc91449037)

[e) Outlier treatment (if required) 20](#_Toc91449038)

[f) Variable transformation (if applicable) 20](#_Toc91449039)

[g) Addition of new variables (if required) 22](#_Toc91449040)

[Business insights from EDA 22](#_Toc91449041)

[a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business 22](#_Toc91449042)

[b) Any business insights using clustering (if applicable) 22](#_Toc91449043)

[c) Any other business insights 23](#_Toc91449044)

[Code 23](#_Toc91449045)

Introduction of the business problem

a) Defining the problem statement

The insurance industry is one of the top 5 industry sectors in India. With recent pandemic situation, the focus on medical insurance has increased immensely. Especially in India, where until 2019 the majority of the population didn’t have a regular medical insurance to cover the medical expenses of their families which can at times be a big amount for middle class salary earning families, a huge section is now looking to buy medical insurance.

In order, to tap into such a huge market potential which suddenly opened in last couple of years, how can the insurance companies expand their business but at the same time ensuring that the risks for the company at the minimum?

b) Need of the study/project

Through this capstone project we would like to solve the above problem for the insurance industry by analyzing and processing a dataset, build a ML model on top of it that would allow us to predict the optimal insurance premium that an individual has to pay. The data has been already collected and has been provided to us. As part of analysis and processing of the dataset, we would explore the data from various angles, perform data cleaning/enrichment before building the model.

c) Understanding business/social opportunity

The outcomes of this project would serve two broad purposes

1. For the business – Predict the optimal cost of premium for their potential customers, keeping the risk as one the key factors
2. For society/customers – Understanding the lifestyle choices, existing health conditions and other factors that may impact the cost of the premium they have to pay to get insured

Data Report

a) Understanding how data was collected in terms of time, frequency and methodology

Following are the understanding about the data collection

1. The maximum year of last admitted is 2018, suggesting the data is collected not earlier than 2019
2. The maximum years of insurance with the company is 8, suggesting the data is collected for at least 8 years. Hence, it would be safe to assume that the data was collected between 2010 and 2018
3. There is a high likelihood that the data was collected as part of medical insurance application process either during a new purchase or during renewal of an existing medical insurance
4. The data is either self-declared by the applicant or captured during medical tests prior to underwriting phase of the application approval process

b) Visual inspection of data (rows, columns, descriptive details)

There are 25000 observations and 24 columns including the identifier column i.e. applicant id. On further inspection, we found that all the features’ values across all the observations except bmi, year\_last\_admitted and smoking\_status. bmi and year\_last\_admitted has null values whereas smoking\_status has unkown observations which need to be addressed during EDA. For now, below is the summary of features from the raw data.

Summary of categorical features:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature name** | **count** | **unique** | **top** | **freq** | **Percent of total** |
| Occupation | 25000 | 3 | Student | 10169 | 40.7% |
| cholesterol\_level | 25000 | 5 | 150 to 175 | 8763 | 35.1% |
| Gender | 25000 | 2 | Male | 16422 | 65.7% |
| smoking\_status | 25000 | 4 | never smoked | 9249 | 37.0% |
| Location | 25000 | 15 | Bangalore | 1742 | 7.0% |
| covered\_by\_any\_other\_company | 25000 | 2 | N | 17418 | 69.7% |
| Alcohol | 25000 | 3 | Rare | 13752 | 55.0% |
| exercise | 25000 | 3 | Moderate | 14638 | 58.6% |

*Table 1: Describing the categorical features in raw data*

Summary of continuous features

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature name** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| applicant\_id | 25000 | 17499.5 | 7217.023 | 5000 | 11250 | 17500 | 23749 | 29999 |
| years\_of\_insurance\_with\_us | 25000 | 4.08904 | 2.606612 | 0 | 2 | 4 | 6 | 8 |
| regular\_checkup\_lasy\_year | 25000 | 0.77368 | 1.199449 | 0 | 0 | 0 | 1 | 5 |
| adventure\_sports | 25000 | 0.08172 | 0.273943 | 0 | 0 | 0 | 0 | 1 |
| visited\_doctor\_last\_1\_year | 25000 | 3.1042 | 1.141663 | 0 | 2 | 3 | 4 | 12 |
| daily\_avg\_steps | 25000 | 5215.8893 | 1053.18 | 2034 | 4543 | 5089 | 5730 | 11255 |
| age | 25000 | 44.91832 | 16.10749 | 16 | 31 | 45 | 59 | 74 |
| heart\_decs\_history | 25000 | 0.05464 | 0.227281 | 0 | 0 | 0 | 0 | 1 |
| other\_major\_decs\_history | 25000 | 0.09816 | 0.297537 | 0 | 0 | 0 | 0 | 1 |
| avg\_glucose\_level | 25000 | 167.53 | 62.72971 | 57 | 113 | 168 | 222 | 277 |
| bmi | 24010 | 31.393328 | 7.876535 | 12 | 26.1 | 30.5 | 35.6 | 101 |
| Year\_last\_admitted | 13119 | 2003.8922 | 7.581521 | 1990 | 1997 | 2004 | 2010 | 2018 |
| weight | 25000 | 71.61048 | 9.325183 | 52 | 64 | 72 | 78 | 96 |
| weight\_change\_in\_last\_one\_year | 25000 | 2.51796 | 1.690335 | 0 | 1 | 3 | 4 | 6 |
| fat\_percentage | 25000 | 28.81228 | 8.632382 | 11 | 21 | 31 | 36 | 42 |
| insurance\_cost | 25000 | 27147.408 | 14323.69 | 2468 | 16042 | 27148 | 37020 | 67870 |

*Table 2: Describing the continuous features in raw data*

Note: Features such as adventure\_sports, heart\_decs\_history, other\_major\_decs\_history are binary flags and values as 0, 1. Though they are categorical features, due to the nature of data capture, the code considers them as continuous feature.

c) Understanding of attributes (variable info, renaming if required)

Below is the understanding of the attributes based on the initial analysis of the variables

|  |  |
| --- | --- |
| **Variable name** | **Variable information** |
| applicant\_id | Identifier of applicants to mark them uniquely in the data |
| years\_of\_insurance\_with\_us | Number of years for which the customer has had existing relationship with the insurance company |
| regular\_checkup\_lasy\_year | Number of times the customers underwent regular medical check-ups during last year. Should be renamed to regular\_checkup\_las~~y~~t\_year |
| adventure\_sports | A flag that indicates customers’ inclination or if they practice any adventure sports which may be hazardous to health or even life threatening, like river rafting, bungee jumping, sky diving etc. |
| Occupation | Type of occupation the customers have |
| visited\_doctor\_last\_1\_year | Number of times the customer consulted a doctor in last one year |
| cholesterol\_level | Total cholesterol levels of the customers, must have been taken during the medical check-up as part of medical insurance application |
| daily\_avg\_steps | Average number of daily steps the customers walks as part of their fitness regime |
| age | Age of the customers in years |
| heart\_decs\_history | A flag that indicated whether the customer has a history of heart related diseases |
| other\_major\_decs\_history | A flag that indicated whether the customer has a history of other major diseases |
| Gender | Gender of the customers |
| avg\_glucose\_level | Average glucose levels of the customers, must have been taken during the medical check-up as part of medical insurance application |
| bmi | BMI of the customers, must have been taken during the medical check-up as part of medical insurance application |
| smoking\_status | Smoking status of the customers at the time of applying for the medical insurance |
| Year\_last\_admitted | The calendar year when the customer was last admitted in hospital with at least one overnight stay |
| Location | City of residence of the applicant |
| weight | Weight in kgs of the customer at the time of applying for the medical insurance |
| covered\_by\_any\_other\_company | A flag that indicated whether the customer has any existing medical insurance with another company |
| Alcohol | Alcohol consumption of the customers at the time of applying for the medical insurance |
| exercise | Overall physical activity regime of the customers at the time of applying for the medical insurance |
| weight\_change\_in\_last\_one\_year | Change in the body weight of the customers in last one year |
| fat\_percentage | Body fat percentage of the customers at the time of applying for the medical insurance |
| insurance\_cost | The cost of premium customer had paid to buy the medical insurance |

*Table 2: Understanding of the attributes*

Exploratory data analysis

a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

Below are the univariate analyses of continuous features available in the dataset

|  |  |  |
| --- | --- | --- |
| # | Distribution plot | Inferences |
| 1 | *Figure 1: Count plot for years of insurance with us* | years\_of\_insurance\_with\_us –   * The distribution of observations across this feature is fairly flat when looked in increasing order of values available * The maximum number is with 3 years but it suddenly dipped the year after (i.e. 2 years of insurance of with us) but then has been increasing since then |
| 2 | *Figure 2: Count plot for regular checkup last year* | regular\_checkup\_last\_year-   * Majority of the customers (only slightly less than 2/3rds of all the customers) have not had any regular medical checkups during last year * About 1,125 customers are either very health conscious or due to existing medical conditions get a regular medical checkup once in a quarter or more |
| 3 | *Figure 3: Count plot for visited doctor last 1 year* | visited\_doctor\_last\_1\_year-   * Approximately about 80% of customers have visited the doctors at least twice in last 1 year |
| 4 | *Figure 4: Count plot for daily average steps split in 10 bins* | daily\_avg\_steps-   * Roughly about 2/3rds of customers log daily average steps between 4000 and 6000 * Less than 1% of customers are able to log the WHO recommended average daily steps of 10000 for healthy lifestyle |
| 5 | *Figure 5: Distribution plot of age of the customers* | age-  The customers are almost evenly distributed in terms of age ranging from 16 years to 74 age |
| 6 | *Figure 6: Distribution plot for average glucose level* | avg\_glucose\_level-   * The average glucose level is also almost evenly distributed across the population. * A healthy glucose level is below 100, suggesting that the customers are not following a very healthy diet regime making them prone to diseases such as type 2 diabetes, hypertension, which may increase their cost of medical insurance |
| 7 | *Figure 7: Distribution plot for BMI of the customers* | bmi-   * The bmi for majority of the population lies between 20 and 40 * The range of bmi index scale between 20 and 40 encapsulates the healthy weight to obese * There appears to some outliers as the kdeplot scales upto 100, which we will further explore later and if needed treat as well |
| 8 | *Figure 8: Count plot for year last admitted* | Year\_last\_admitted-   * The observations appear to be evenly distributed between years 1993 and 2014. * It should also be noted here that almost 50% of the observations in this feature are missing * The missing values can be a legitimate case as it is not always necessary that everyone in the population is expected to be admitted in hospital at some point of time |
| 9 | *Figure 9: Distribution plot for weight of the customers* | weight-   * The observations are not normally distributed with a couple of peaks on the distribution plot * Majority of the population has body weight between 60 and 80 kgs |
| 10 | ` *Figure 10: Count for weight change in last one year* | weight\_change\_in\_last\_one\_year-   * Most of the customers have shown low to moderate weight loss in last one year * Within the above majority of the customers have lost between 3-4 kgs of body weight in last one year |
| 11 | *Figure 11: Distribution plot for fat percentage* | fat\_percentage-   * The observations are not normally distributed with many peaks on the distribution plot * Majority of observations have body fat percent between 20-35% which much higher than healthy norms irrespective of age and gender |

*Table 4: Univariate analysis of continuous features*

Below are the univariate analyses of categorical features available in the dataset

|  |  |  |
| --- | --- | --- |
| # | Count plot | Inferences |
| 1 | *Figure 12: Count plot for adventure sports* | adventure\_sports-   * Almost 90% of the customers do not play any adventure sports which can be a threat to health/life of the customers * For the remaining customers the expected cost of medical insurance would be on a higher side as their risk profile will be higher |
| 2 | *Figure 13: Count plot for Occupation* | Occupation-   * Customers across Student and Business are almost equal * Salaried class customers are roughly 1/5th of all the customers willing to purchase/renew the medical insurance |
| 3 | *Figure 14: Count plot for cholesterol level* | cholesterol\_level-   * Majority of the customers have good control on their cholesterol levels which should be below 200 * Customers with higher cholesterol levels are at higher risk of developing heart related diseases |
| 4 | *Figure 15: Count plot for heart diseases history* | heart\_decs\_history-   * A very small portion of all the customers have existing heart related diseases which puts them in high risk category for the insurance companies |
| 5 | *Figure 16: Count plot for other major diseases history* | other\_major\_decs\_history-   * There is also a small portion of customers who have other comorbidities which can also be considered as higher risk customers for the insurance company |
| 6 | *Figure 17: Count plot for Gender* | Gender-   * The male customers are roughly twice as much as the female customers |
| 7 | *Figure 18: Count plot for smoking status* | smoking\_status-   * Roughly around 1/3rd of the customers either smoked in the past or are currently smoking * The customers who are currently smoking are definitely high-risk customers as their chances of developing heart or lungs related diseases are higher * For an appreciable section of customers, the smoking status is unknown, we will have to treat this information in later stages prior to model building |
| 8 | *Figure 19: Count plot for Location* | Location-   * The customers are evenly distributed across the Indian cities * There is no specific trend which can help us make inferences through univariate analysis |
| 9 | *Figure 20: Count plot for covered by other company* | covered\_by\_any\_other\_company-   * Roughly around 30% of the customers are covered by other insurance companies |
| 10 | *Figure 21: Count plot for Alcohol* | Alcohol-   * Slightly above 1/3rd of the customers doesn’t consume alcohol * Customers with daily alcohol consumption are at higher risk of developing liver related diseases and other comorbidities * Customers with rare alcohol consumption habit are also at higher risk if their other life style choices, such as no exercise, smoking etc. in combination may create health related problems |
| 11 | *Figure 22: Count plot for exercise* | Exercise-   * Customers with no physical exercise in their routine are at higher medical risk and hence the cost to insurance will be on higher side |

*Table 5: Univariate analysis of categorical features*

b) Bivariate analysis (relationship between different variables, correlations)

Weight

|  |  |  |
| --- | --- | --- |
| # | Count plot | Inferences |
| 1 | *Figure 23: Box plot between weight and smoking status* | smoking\_status-   * There isn’t any evidence of strong correlation between smoking status and weight of the customers |
| 2 | *Figure 24: Box plot between weight and alcohol* | Alcohol-   * There isn’t any evidence of strong correlation between Alcohol and weight of the customers |
| 3 | *Figure 25: Box plot between weight and exercise* | Exercise-   * There isn’t any evidence of strong correlation between exercise and weight of the customers |
| 4 | *Figure 26: Box plot between weight and Occupation* | Occupation-   * There isn’t any evidence of strong correlation between Occupation and weight of the customers |
| 5 | *Figure 27: Box plot between weight and cholesterol level* | cholesterol\_level-   * There isn’t any evidence of strong correlation between Cholesterol levels and weight of the customers |
| 6 | *Figure 28: Box plot between weight and Gender* | Gender   * There isn’t any evidence of strong correlation between Gender and weight of the customers |
| 7 | *Figure 29: Box plot between weight and heart disease history* | heart\_decs\_history   * There isn’t any evidence of strong correlation whether the customers had a history of heart diseases and their weight |
| 8 | *Figure 30: Box plot between weight and other major heart disease history* | other\_major\_decs\_history-   * There isn’t any evidence of strong correlation whether the customers had a history of any other major diseases and their weight |
| 9 | *Figure 31: Box plot between weight and covered by other company* | covered\_by\_any\_other\_company-   * There isn’t any evidence of strong correlation whether the customers are covered by other companies and their weight |
| 10 | *Figure 32: Box plot between weight and regular check last year* | regular\_checkup\_lasy\_year-   * Customers who undergo regular health checkups tend to have lesser weight as compared to customers who take less often or no health checkups |

*Table 6: Bivariate analysis of Weight with categorical features*

Average glucose level

|  |  |  |
| --- | --- | --- |
| # | Count plot | Inferences |
| 1 | *Figure 33: Box plot between average glucose level and smoking status* | smoking\_status-   * There isn’t any evidence of strong correlation between smoking status and average glucose levels of the customers |
| 2 | *Figure 34: Box plot between average glucose level and alcohol* | Alcohol   * There isn’t any evidence of strong correlation between alcohol consumption frequency and average glucose levels of the customers |
| 3 | *Figure 35: Box plot between average glucose level and exercise* | Exercise-   * There isn’t any evidence of strong correlation between exercise frequency and average glucose levels of the customers |
| 4 | *Figure 36: Box plot between average glucose level and Occupation* | Occupation-   * There isn’t any evidence of strong correlation between Occupation and average glucose levels of the customers |
| 5 | *Figure 37: Box plot between average glucose level and cholesterol level* | cholesterol\_level-   * There isn’t any evidence of strong correlation between cholesterol levels and average glucose levels of the customers |
| 6 | *Figure 38: Box plot between average glucose level and Gender* | Gender-   * There isn’t any evidence of strong correlation between Gender and average glucose levels of the customers |
| 7 | *Figure 39: Box plot between average glucose level and heart disease history* | heart\_decs\_history-   * There isn’t any evidence of strong correlation whether the customers had a history of heart diseases and their average glucose levels |
| 8 | *Figure 40: Box plot between average glucose level and other major disease history* | other\_major\_decs\_history-   * There isn’t any evidence of strong correlation whether the customers had a history of any other comorbidities and their average glucose levels |
| 9 | *Figure 41: Box plot between average glucose level and covered by other company* | covered\_by\_any\_other\_company-   * There isn’t any evidence of strong correlation whether the customers are covered by other insurance companies and their average glucose levels |
| 10 | *Figure 42: Box plot between average glucose level and regular check last year* | regular\_checkup\_lasy\_year-   * There isn’t any evidence of strong correlation between frequency of health checkups and average glucose levels of the customers |

*Table 7: Bivariate analysis of Average glucose levels with categorical features*

BMI

|  |  |  |
| --- | --- | --- |
| # | Count plot | Inferences |
| 1 | *Figure 43: Box plot between BMI and smoking status* | smoking\_status-   * Customers who have never smoked or had unknown smoking status have lower BMI when compared to customers who currently smoke or had smoked in the past |
| 2 | *Figure 44: Box plot between BMI and alcohol* | Alcohol   * There isn’t any evidence of strong correlation between alcohol consumption frequency and BMI of the customers |
| 3 | *Figure 45: Box plot between BMI and exercise* | Exercise-   * There isn’t any evidence of strong correlation between exercise frequency and BMI of the customers |
| 4 | *Figure 46: Box plot between BMI and Occupation* | Occupation-   * There isn’t any evidence of strong correlation between Occupation and BMI of the customers |
| 5 | *Figure 47: Box plot between BMI and cholesterol level* | cholesterol\_level-   * There isn’t any evidence of strong correlation between cholesterol levels and BMI of the customers |
| 6 | *Figure 48: Box plot between BMI and Gender* | Gender-   * Females have lower BMI when compared to male customers |
| 7 | *Figure 49: Box plot between BMI and heart disease history* | heart\_decs\_history-   * Customers with an history of heart diseases have on an average slightly higher BMI |
| 8 | *Figure 50: Box plot between BMI and other major disease history* | other\_major\_decs\_history-   * Customers with an history of any other comorbidities have on an average slightly higher BMI |
| 9 | *Figure 51: Box plot between BMI and covered by other company* | covered\_by\_any\_other\_company-   * There isn’t any evidence of strong correlation whether the customers are covered by other insurance companies and their BMI |
| 10 | *Figure 52: Box plot between BMI and regular check last year* | regular\_checkup\_lasy\_year-   * Customers who undergo regular health checkups tend to have lower outliers in the BMI as compared to customers who take less often or no health checkups |

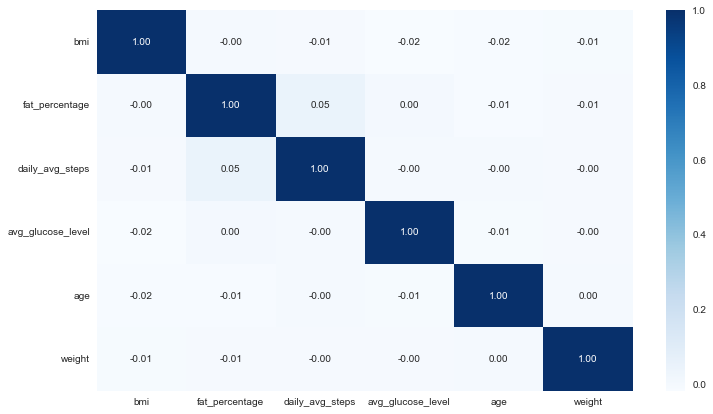
*Table 8: Bivariate analysis of BMI with categorical features*

Fat percentage

|  |  |  |
| --- | --- | --- |
| # | Count plot | Inferences |
| 1 | *Figure 53: Box plot between fat percentage and smoking status* | smoking\_status-   * There isn’t any evidence of strong correlation between smoking status and fat percentage of the customers |
| 2 | *Figure 54: Box plot between fat percentage and alcohol* | Alcohol   * Customers who don’t consume alcohol have lower fat percentage as compared to customers who consume alcohols |
| 3 | *Figure 55: Box plot between fat percentage and exercise* | Exercise-   * There isn’t any evidence of strong correlation between exercise frequency and fat percentage of the customers |
| 4 | *Figure 56: Box plot between fat percentage and Occupation* | Occupation-   * Salaried customers have lower fat percentage when compared to Students or Business class customers |
| 5 | *Figure 57: Box plot between fat percentage and cholesterol level* | cholesterol\_level-   * There seems to be strong correlation between cholesterol levels and fat percentage of the customers. Lower cholesterol customers tend to have lower fat percentage |
| 6 | *Figure 58: Box plot between fat percentage and Gender* | Gender-   * There isn’t any evidence of strong correlation between Gender and fat percentage of the customers |
| 7 | *Figure 59: Box plot between fat percentage and heart disease history* | heart\_decs\_history-   * There isn’t any evidence of strong correlation whether the customers had a history of heart diseases and their fat percentages |
| 8 | *Figure 60: Box plot between fat percentage and other major disease history* | other\_major\_decs\_history-   * There isn’t any evidence of strong correlation whether the customers had a history of any other comorbidities and their fat percentages |
| 9 | *Figure 61: Box plot between fat percentage and covered by other company* | covered\_by\_any\_other\_company-   * There isn’t any evidence of strong correlation whether the customers are covered by other insurance companies and their fat percentages |
| 10 | *Figure 62: Box plot between fat percentage and regular check last year* | regular\_checkup\_lasy\_year-   * The average fat percentage tends to be lower for customers who undergo medical checkups on a regular frequency |

*Table 9: Bivariate analysis of Body fat percentage with categorical features*

Bivariate analysis between continuous features



*Figure 63: Correlation heat map between the continuous features*

It is evident from the above heat map of correlation plot amongst the continuous features, BMI, fat\_percentage, daily\_avg\_steps, avg\_glucose\_level, age & weight, there isn’t any strong correlation between any of the features.

c) Removal of unwanted variables (if applicable)

Post EDA, while preparing the data for model building some of the categorical and/or the features with object data type will have to be replaced with their numerical counterpart, either through binary encoding or in case when the features qualify as ordinal categorical variables then with resp. numerical evaluations. We will cover this more in f) Variable transformation section of this module.

d) Missing Value treatment (if applicable)

There are three features where we have missing or unknown features. The treatment of missing values for each of these features should be done keeping the context of the data.

BMI – out of 25000 observations, there are 990 observations with missing values of BMI. Ideally, BMI is derived using body weight and height. We do have body weight but there is no height provided in the data. Alternatively, BMI can be reverse calculated using fat percentage of the body. We do have fat percentage provided in the data but there is no correlation between the BMI and fat percentage as seen in the previous section. Hence, it will be unsafe to impute the missing values by reverse calculating it using fat percentage. We will impute the data by usual methodology of utilized for continuous features of mean or median. But since, there are outliers in the data it would be safe to impute the missing values using median value.

Year\_last\_admitted – out of 25000 observations, there are 11881 observations with missing values of year\_last\_admitted. This is roughly about 47.5% of data that has missing values. In general, with such high percentage of missing values in a feature makes it unusual for any sorts of model building. But if we look at the context, not all individuals do need to get admitted in hospitals with overnight stay. In US, per a government website only 7.9% of persons need an overnight stay at hospitals[[Link](https://www.cdc.gov/nchs/fastats/hospital.htm)]. So, it would be safe to default all the missing values to a starting point to bring the data across all observations on same scale. If we impute year 1900 against all the missing values and later standardize(scale) the data then all the observations of this feature will be homogenous.

Smoking\_status – out of 25000 observations, there are 7555 observations with status as Unknown. Through bivariate analysis we saw that there isn’t much correlation with any of the continuous features except BMI where for Unknown category the BMI was on lower side similar to never smoked category. Since, never smoked category is also the most frequently occurring the category for this feature, it will be safe to impute the Unknown with the mode of the remaining observations i.e. never smoked.

e) Outlier treatment (if required)

Out of 11 continuous features, below are the percentages of outliers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Name** | **Percentage outliers** | **Median** | **Minimum** | **Maximum** |
| years\_of\_insurance\_with\_us | 0.00% | 4 | 0 | 8 |
| regular\_checkup\_lasy\_year | 11.77% | 0 | 0 | 5 |
| visited\_doctor\_last\_1\_year | 12.16% | 3 | 0 | 12 |
| daily\_avg\_steps | 15.96% | 5089 | 2034 | 11255 |
| age | 15.96% | 45 | 16 | 74 |
| avg\_glucose\_level | 15.96% | 168 | 57 | 277 |
| bmi | 18.46% | 30.5 | 12.3 | 100.6 |
| Year\_last\_admitted | 18.46% | 1992 | 1900 | 2018 |
| weight | 18.46% | 72 | 52 | 96 |
| weight\_change\_in\_last\_one\_year | 18.46% | 3 | 0 | 6 |
| fat\_percentage | 18.46% | 31 | 11 | 42 |

*Table 10: Describing the outliers, median, minimum and maximum of continuous features*

We can see that except for BMI, the minimums and maximums of all the other variables are within acceptable range. Hence, only BMI should be treated for its outliers. Post outlier treatment the statistics for the feature looks much more acceptable especially towards the upper range, as can be seen below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Name** | **Percentage outliers** | **Median** | **Minimum** | **Maximum** |
| bmi | 0.00% | 30.5 | 12.8 | 48.8 |

*Table 11: Outliers removed after treatment of data under BMI*

f) Variable transformation (if applicable)

Data needs to transformed prior to model building due two primary reasons.

1. Presence of non-numerical categorical variables: These categorical variables need to transformed in numeric fields considering whether they are ordinal in nature or are random values
2. Scale of data: The data across fields are in different range of scales. For example, year\_last\_admitted is in range between 1900 and 2018 whereas the other fields are less than 100. Since, some of the models are distance based, having data in different scales will bias the model training in favor of variables with larger scale making the predictions inaccurate

**Transforming non-numeric categorical variables**

We will not be using Binary encoding as it would increase the number of features in the data reducing the accuracy of the models at later stages. Instead for each feature based on the type of data it has we will pick an appropriate method to replace the data with numeric value

|  |  |
| --- | --- |
| **Feature name** | **Transformation strategy** |
| Occupation | There is no order in the data and hence, it would safe to assume that the data is random. We would replace the existing values with 1000, 1001, 1002 resp. The numbers are unique yet will not bias the data due to its ordinality as the scale of 1000 vs 1 is very wide |
| cholesterol\_level | The data is ordinal in nature where 125 to 150 is best and 225 to 250 is worst. Hence, the values were replaced in decreasing order from 5 to 1, where 5 was assigned to best value and 1 to the worst value |
| Gender | There is no order in the data and hence, it would safe to assume that the data is random. We would replace the existing values with 1000, 1001 resp. The numbers are unique yet will not bias the data due to its ordinality as the scale of 1000 vs 1 is very wide |
| smoking\_status | The data is ordinal in nature where ‘never smoked’ is best and ‘smokes’ is worst. Hence, the values were replaced in decreasing order from 3 to 1, where 3 was assigned to best value and 1 to the worst value |
| Location | There is no order in the data and hence, it would safe to assume that the data is random. We would replace the existing values with 1000, 1014 resp. The numbers are unique yet will not bias the data due to its ordinality as the scale of 1000 vs 1 is very wide |
| covered\_by\_any\_other\_company | N was replaced with 0 and Y with 1 |
| Alcohol | The data is ordinal in nature where ‘Daily’ is best and smokes is ‘No’. Hence, the values were replaced in decreasing order from 3 to 1, where 3 was assigned to best value and 1 to the worst value |
| exercise | The data is ordinal in nature where ‘Extreme’ is best and ‘No’ is worst. Hence, the values were replaced in decreasing order from 3 to 1, where 3 was assigned to best value and 1 to the worst value |

*Table 12: Strategy of replacing non-numerical categorical features with numbers for data preparation*

**Scaling of the data**

After conversion of all the non-numerical categorical variables, the original variables will be dropped and only numerical features are retained in the dataframe. The data now needs to be scaled. Below are the first five records from the data before and after scaling.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| First 5 records before scaling, but with non-numeric variables replaced with numeric values | | | | | | | | | | | | | | | | | | | | | | |
| years\_of\_insurance\_with\_us | regular\_checkup\_lasy\_year | adventure\_sports | visited\_doctor\_last\_1\_year | daily\_avg\_steps | age | heart\_decs\_history | other\_major\_decs\_history | avg\_glucose\_level | bmi | Year\_last\_admitted | weight | weight\_change\_in\_last\_one\_year | fat\_percentage | cholesterol\_level\_num | smoking\_status\_num | Alcohol\_num | exercise\_num | covered\_by\_any\_other\_company\_num | Location\_num | Occupation\_num | Gender\_num |
| 3 | 1 | 1 | 2 | 4866 | 28 | 1 | 0 | 97 | 31.2 | 1900 | 67 | 1 | 25 | 5 | 3 | 2 | 2 | 0 | 1000 | 1000 | 1000 |
| 0 | 0 | 0 | 4 | 6411 | 50 | 0 | 0 | 212 | 34.2 | 1900 | 58 | 3 | 27 | 4 | 2 | 2 | 2 | 0 | 1001 | 1001 | 1000 |
| 1 | 0 | 0 | 4 | 4509 | 68 | 0 | 0 | 166 | 40.4 | 1900 | 73 | 0 | 32 | 2 | 2 | 1 | 3 | 0 | 1001 | 1002 | 1001 |
| 7 | 4 | 0 | 2 | 6214 | 51 | 0 | 0 | 109 | 22.9 | 1900 | 71 | 3 | 37 | 3 | 3 | 2 | 1 | 1 | 1000 | 1002 | 1001 |
| 3 | 1 | 0 | 2 | 4938 | 44 | 0 | 1 | 118 | 26.5 | 2004 | 74 | 0 | 34 | 4 | 3 | 3 | 3 | 0 | 1002 | 1001 | 1000 |

*Table 13: First 5 records before scaling, but with non-numeric variables replaced with numeric values*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| First 5 records after scaling | | | | | | | | | | | | | | | | | | | | | | |
| years\_of\_insurance\_with\_us | regular\_checkup\_lasy\_year | adventure\_sports | visited\_doctor\_last\_1\_year | daily\_avg\_steps | age | heart\_decs\_history | other\_major\_decs\_history | avg\_glucose\_level | bmi | Year\_last\_admitted | weight | weight\_change\_in\_last\_one\_year | fat\_percentage | cholesterol\_level\_num | smoking\_status\_num | Alcohol\_num | exercise\_num | covered\_by\_any\_other\_company\_num | Location\_num | Occupation\_num | Gender\_num |
| -0.42 | 0.19 | 3.35 | -0.97 | -0.33 | -1.05 | 4.16 | -0.33 | -1.12 | 0.00 | -1.04 | -0.49 | -0.90 | -0.44 | 1.00 | 0.65 | -0.37 | -0.01 | -0.66 | -1.61 | -1.63 | -0.72 |
| -1.57 | -0.65 | -0.30 | 0.78 | 1.13 | 0.32 | -0.24 | -0.33 | 0.71 | 0.42 | -1.04 | -1.46 | 0.29 | -0.21 | 0.21 | -0.69 | -0.37 | -0.01 | -0.66 | -1.38 | -0.28 | -0.72 |
| -1.19 | -0.65 | -0.30 | 0.78 | -0.67 | 1.43 | -0.24 | -0.33 | -0.02 | 1.29 | -1.04 | 0.15 | -1.49 | 0.37 | -1.37 | -0.69 | -1.96 | 1.55 | -0.66 | -1.38 | 1.07 | 1.38 |
| 1.12 | 2.69 | -0.30 | -0.97 | 0.95 | 0.38 | -0.24 | -0.33 | -0.93 | -1.16 | -1.04 | -0.07 | 0.29 | 0.95 | -0.58 | 0.65 | -0.37 | -1.56 | 1.52 | -1.61 | 1.07 | 1.38 |
| -0.42 | 0.19 | -0.30 | -0.97 | -0.26 | -0.06 | -0.24 | 3.03 | -0.79 | -0.66 | 0.95 | 0.26 | -1.49 | 0.60 | 0.21 | 0.65 | 1.22 | 1.55 | -0.66 | -1.14 | -0.28 | -0.72 |

*Table 14: First 5 records after scaling*

g) Addition of new variables (if required)

The non-numeric variables were replaced with numeric counter parts. No new variables were featured out of existing data provided. Ideally, additional information such as hereditary medical conditions, educational qualification, more discrete occupation details, etc.

Business insights from EDA

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business

For certain features the data is unbalanced. For example, BMI and weights are not normally distributed. Also, categorical features such as Gender, Alcohol and Exercise are unbalanced. There are under or over sampling techniques to balance the data when the need is to bias the models in favor of a certain feature. But nothing can beat the accuracy of model that is trained on good volume data. In order to offset the bias due to unbalanced data, increasing the number of observations strengthens the predictions and makes the model more generic. A model trained on a big dataset tends to perform well for longer duration and much diverse scenarios than compared to a model trained on a small dataset.

Coming back to the data at hand, the data imbalance is unavoidable and is very much the representation of what one should expect with any data of this sorts, if a smaller sample size is picked. This is primarily because of social structure, a wide variety of preferences and habits people develop or drop at different events in their life etc. As earlier discussed, bigger datasets both in length and breadth will help offset the bias a model may develop due to imbalance in the data. While collecting the data business should include additional features as below.

* Existing medical conditions of blood relationships, like father, mother, siblings, grandparents etc.
* Highest educational qualifications
* Annual household income
* Occupation details to determine the type of work, whether sedentary or field work
* Additional details on type of existing comorbidities
* Whether currently on any medication

b) Any business insights using clustering (if applicable)

The scaled data created in Variable transformation section, was further to cluster the data and segment the customers. Several iterations for different count of clusters, via K-Means clustering, were performed to find the best silhouette scores, as the scree plot for weighted sums of squares wasn’t conclusive to find the optimal number of clusters. Comparing the silhouette scores, it was found that 3 clusters would be the most efficient way of splitting the data.

|  |  |  |  |
| --- | --- | --- | --- |
| # clusters | Silhouette score | Min. Silhouette sample | *Figure 64: Scree plot* |
| 3 | 0.076975 | -0.00792 |
| 4 | 0.067023 | -0.02425 |
| 5 | 0.062753 | -0.03804 |
| 6 | 0.060842 | -0.02132 |
| 7 | 0.062669 | -0.04702 |
| 9 | 0.062077 | -0.02748 |
| 10 | 0.0565 | -0.06052 |

*Table 15: Comparison of silhouette scores*

Through hierarchical clustering there wasn’t a good cut off point that would help determine the optimal number clusters for segmentation.

Once the observations were labeled using the K Means clustering, the comparison of the cluster profiles revealed that except years\_of\_insurance\_with\_us and regular\_checkup\_last\_year and some extent with the insurance\_cost, all the remaining features showed very slight deviation across the three clusters. For business this could mean the following.

1. The features available in the data appear incapable of directly impacting the cost of the insurance alone on their own
2. There are multiple factors that come into play to determine the final cost of insurance for each customer
3. The features are well distributed and do not show any trend with respect to the how customers behave
4. Being able to segment the customers would have gone a long way in drafting specific products for each segment

c) Any other business insights

Following are the additional business insights based on the data we have.

1. Knowing the customer well is the key to decide the risk factor for the company but at the same offer the best product for their customers
2. Understanding the risk profile of the customers, their preferences, ensuring once a customer always a customer is key to the success of the business
3. Offering additional services, like membership of fitness clubs, tie ups with pharmacies for discounts on medicines, or consultation with specialists would not only attract more customers but would help decrease the risk profile of the customers in longer run

Part 1 Code 

Github repository to the code & business report: <https://github.com/sarang-manohar/hogwarts/tree/main/academia/capStone>

****Model building and interpretation****

a. Build various models and test your predictive model against the test set using various appropriate performance metrics

The intention of the model is to predict insurance cost based on the independent variables made available. In previous sections, the data was processed and treated to take care of the null values, outliers etc. Finally, the data was scaled(normalized) to suppress any bias that the data may cause. The target variable is continuous and hence the regression models need to be trained for the use case at hand. The base version of various models was trained and predictions were made to validate the scores against the both train and test data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | model | R2\_train | R2\_test | RMSE\_train | RMSE\_test | MAPE\_train | MAPE\_test |
| 1 | LinearRegression() | 0.94 | 0.95 | 3367.8 | 3343.5 | 15.5% | 14.8% |
| 2 | Ridge() | 0.94 | 0.95 | 3367.8 | 3343.5 | 15.5% | 14.8% |
| 3 | Lasso() | 0.94 | 0.95 | 3367.8 | 3343.4 | 15.5% | 14.8% |
| 4 | BayesianRidge() | 0.94 | 0.95 | 3367.8 | 3343.5 | 15.5% | 14.8% |
| 5 | ARDRegression() | 0.94 | 0.95 | 3368.4 | 3343.2 | 15.5% | 14.8% |
| 6 | SGDRegressor() | 0.94 | 0.95 | 3376.7 | 3356.2 | 15.4% | 14.8% |
| 7 | PassiveAggressiveRegressor() | 0.94 | 0.95 | 3374.1 | 3352.7 | 15.4% | 14.7% |
| 8 | ElasticNet() | 0.84 | 0.84 | 5708.6 | 5827.1 | 27.7% | 27.7% |
| 9 | DecisionTreeRegressor() | 1.00 | 0.91 | 0.0 | 4255.1 | 0.0% | 15.3% |
| 10 | ExtraTreesRegressor() | 1.00 | 0.95 | 0.1 | 3148.3 | 0.0% | 11.8% |
| 11 | RandomForestRegressor() | 0.99 | 0.96 | 1138.0 | 3067.6 | 4.3% | 11.5% |
| 12 | BaggingRegressor() | 0.99 | 0.95 | 1338.5 | 3216.0 | 4.7% | 12.0% |
| 13 | GradientBoostingRegressor() | 0.96 | 0.96 | 2929.6 | 2992.6 | 11.4% | 11.4% |
| 14 | AdaBoostRegressor() | 0.95 | 0.95 | 3280.3 | 3302.5 | 15.7% | 15.2% |
| 15 | XGBRegressor() | 0.98 | 0.95 | 2116.1 | 3104.7 | 8.2% | 11.9% |

1. Linear regression
2. Ridge regression
3. Lasso regression
4. BayesianRidge
5. ARDRegression
6. SRDRegression
7. PassiveAggressiveRegression
8. Elasticnet regression
9. DecisionTreeRegressor
10. ExtraTreesRegressor
11. Random forest regression
12. BaggingRegressor
13. Boosting - GradientBoost
14. Boosting - AdaBoost
15. Boosting – XGBoost
16. Neural Network regression
17. Stepwise regression

b. Interpretation of the model(s)

****Model Tuning and business implication****

a. Ensemble modelling, wherever applicable

b. Any other model tuning measures

c. Interpretation of the most optimum model and its implication on the business