EAS 508
STATISTICAL LEARNING AND DATA MINING-I

FINAL PROJECT REPORT

PROJECT TITLE

MEDICAL INSURANCE PREMIUM PREDICTION

TEAM 9:

Meghna Pal-50475251

Jaipriya Gunasekar-50475836

Mohana Priya Gandhiram-50478315

Roshni Balasubramanian-50483779

TABLE OF CONTENTS:

INTRODUCTION:

Health is wealth. Staying healthy is the most important aspect of human life cycle. Now healthcare comes into the picture. Healthcare helps to prevent diseases and improve the quality of life. Health insurance works based on the amount regularly paid by the customer. The amount collected can vary depending on the state of health and the need for treatment. The goal of this project is to find the optimal amount that people will pay based on certain characteristics. The price per customer varies based on the type of medical condition, the need for treatment, the frequency of visits and more. Again, it depends on several aspects such as previous lifestyle, physical condition, health status or genetics. Some external factors that cannot be controlled such as pollution, water quality etc. may depend on location and others. The amount collected should vary based on above factors, as clients who are less likely to need health care should be encouraged to pay lower amounts than those who are at higher risk of receiving medical care. The amount collected from the patients who are in high peril should not be so large that it loses them, and the patients who are in low peril should not be so low as to not have enough funds to pay for the treatments required, while also keeping under consideration the profits of the company. Thus, based on certain characteristics, the optimal amount that each person pays must be found. In this project, we try to find this optimal amount using models, considering the values of the variables. We have data containing information on 25,000 patients for several years. Using the data models, we can predict the insurance cost that has to be paid by the customer. The predictive models will help to estimate the present and future patient insurance cost.

DATASET DESCRIPTION:

The data available to us has been collected over a period of approximately 30 years (currently 2022). The exact method of collection is unknown, but we can speculate that either the patient was recently examined, or the data was transferred from a book to a digital archive. Each new client requires information to be filled in as a new record for all variables, and each time a patient goes for treatment or testing, hospital administrators inform us of the patient's condition when applying for insurance. In this data we have 25000 records. The data is collected based on the 24 attributes. In the given table, each attribute name and the description of that attribute is given.

ATTRIBUTE NAME	DESCRIPTION OF THE ATTRIBUTE
applicant_id	Applicant unique ID
years_of_insurance_with_us	Number of years the customer is taking policy from the same company
regular_checkup_lasy_year:	Number of times the customer has done the regular checkup in last one year
adventure_sports	Customers involved in adventure sports
Occupation	Occupation of the customer
visited_doctor_last_1_year	Number of times customer has visited doctor in last one year
cholesterol_level	Level of cholesterol while applying for insurance
daily_avg_steps	Average steps the customer walks
age	Customer age
heart_decs_history	History of heart diseases
insurance_cost	Total insurance cost
other_major_decs_history	History of any other diseases apart from heart disease
Gender	Customer's gender
avg_glucose_level	Glucose level of the customer while applying the insurance
Bmi	BMI of the customer
smoking_status	Smoking of the customer
Year_last_admitted	Customer admitted in the hospital for the last year
Location	Location of the hospital
weight	Weight of the customer
covered_by_any_other_company	Customer whose insurance is covered by another company
Alcohol	Alcohol consumption status of the customer
exercise	Regular exercise status of the customer
weight_change_in_last_one_year	Variation in customer weight
fat_percentage	Fat percentage while applying the insurance

RENAMING OF COLUMN NAMES:

The given column names are long. So, we are reducing the names to the shorter alternatives.

ATTRIBUTE NAME	SHORTER VERSION OF THE ATTRIBUTE
applicant_id	NOT CHANGED
years_of_insurance_with_us	YOI_us
regular_checkup_lasy_year:	Reg_checkup_lst_yr
adventure_sports	Adv_sports
Occupation	NOT CHANGED
visited_doctor_last_1_year	Dr_visit_lst_yr
cholesterol_level	NOT CHANGED
daily_avg_steps	NOT CHANGED
age	NOT CHANGED
heart_decs_history	Hrt_disease_hist
insurance_cost	NOT CHANGED
other_major_decs_history	Other_mjr_disease_hist
Gender	gender
avg_glucose_level	Avg_glucose_lvl
Bmi	BMI
smoking_status	Smoking_status
Year_last_admitted	Year_last_admtd
Location	Location
weight	weight
covered_by_any_other_company	Other_company_cover
Alcohol	Alcohol
exercise	Exercise
weight_change_in_last_one_year	Wt_chng_lst_yr
insurance_cost	Insurance_cst
fat_percentage	Fat_prcnt

CHECK FOR DATATYPE IN DATAFRAME

YOI_us**,** Reg_checkup_lst_yr**,** Adv_sports, Dr_visit_lst_yr, Daily_avg_steps, Age, Hrt_disease_hist, Othr_mjr_disease_hist, Avg_glucose_lvl, Weight, Wt_chng_lst_yr, Fat_prcnt, Insurance_cst has the datatype **int64.** Occupation, Cholesterol level, Gender, Smoking_status, Location, Other_company_cover, Alcohol, Exercise has the datatype **Object**.There are 24 attributes in that the datatype of the attributes are 2 are float, 8 are object and 14 are integers.

EXPLORATORY DATA ANALYSIS:

DROPPING ATTRIBUTES:

There are 24 attributes in the data, and prediction becomes a tedious task as the dimensionality increases significantly. To solve this problem, you can remove columns based on their relevance and importance to the situation at hand. The application_id Is a unique attribute for each of the patient, hence it is of no use for the prediction of the amount. Therefore, the variable is dropped from the dataset. All the other attributes can be kept since they are important factor. Attributes like adv_sports can be conflicting variables. This indicates that patients engaged in adventure sports may be healthy, but they also increase the risk of injury, requiring medical attention and insurance coverage. In such overlapping situations, it becomes difficult to omit other attributes as there are non-overlapping factors that may play a greater role in generating the best prediction. That said, there are some attributes that aren't currently cleared, but that may not be a significant factor, and such columns will be cleared as needed. Insurance_cst is a target variable, so it cannot be deleted and must be included in the data. In the below table we can find the count, mean, standard deviation, minimum of each of the attribute. In the below table it indicates that there are two missing values in 2 attributes. The data is right skewed. All the variables are given equal weightage, since the range of the variables differ significantly. Attributes BMI and Year_last_admtd has missing values. BMI has 990 missing values, which is 3.96% of the entire data and Year_last_admtd has 11881 missing values, which accounts for 47.52% of the entire data.

	Count	Mean	Std. Dev.	Min	25%	50%	75%	Max
YOI_us	25000	4.08904	2.606612	0	2	4	6	8
Reg_checkup_lst_yr	25000	0.77368	1.199449	0	0	0	1	5
Adv_sports	25000	0.08172	0.273943	0	0	0	0	1
Dr_visit_lst_yr	25000	3.1042	1.141663	0	2	3	4	12
Daily_avg_steps	25000	5215.88932	1053.17975	2034	4543	5089	5730	11255
Age	25000	44.91832	16.107492	16	31	45	59	74
Hrt_disease_hist	25000	0.05464	0.227281	0	0	0	0	1
Othr_mjr_disease_hist	25000	0.09816	0.297537	0	0	0	0	1
Avg_glucose_lvl	25000	167.53	62.729712	57	113	168	222	277
BMI	24010	31.393328	7.876535	12.3	26.1	30.5	35.6	100.6
Year_last_admtd	13119	2003.89222	7.581521	1990	1997	2004	2010	2018
Weight	25000	71.61048	9.325183	52	64	72	78	96
Wt_chng_lst_yr	25000	2.51796	1.690335	0	1	3	4	6
Fat_prcnt	25000	28.81228	8.632382	11	21	31	36	42
Insurance_cst	25000	27147.4077	14323.6918	2468	16042	27148	37020	67870

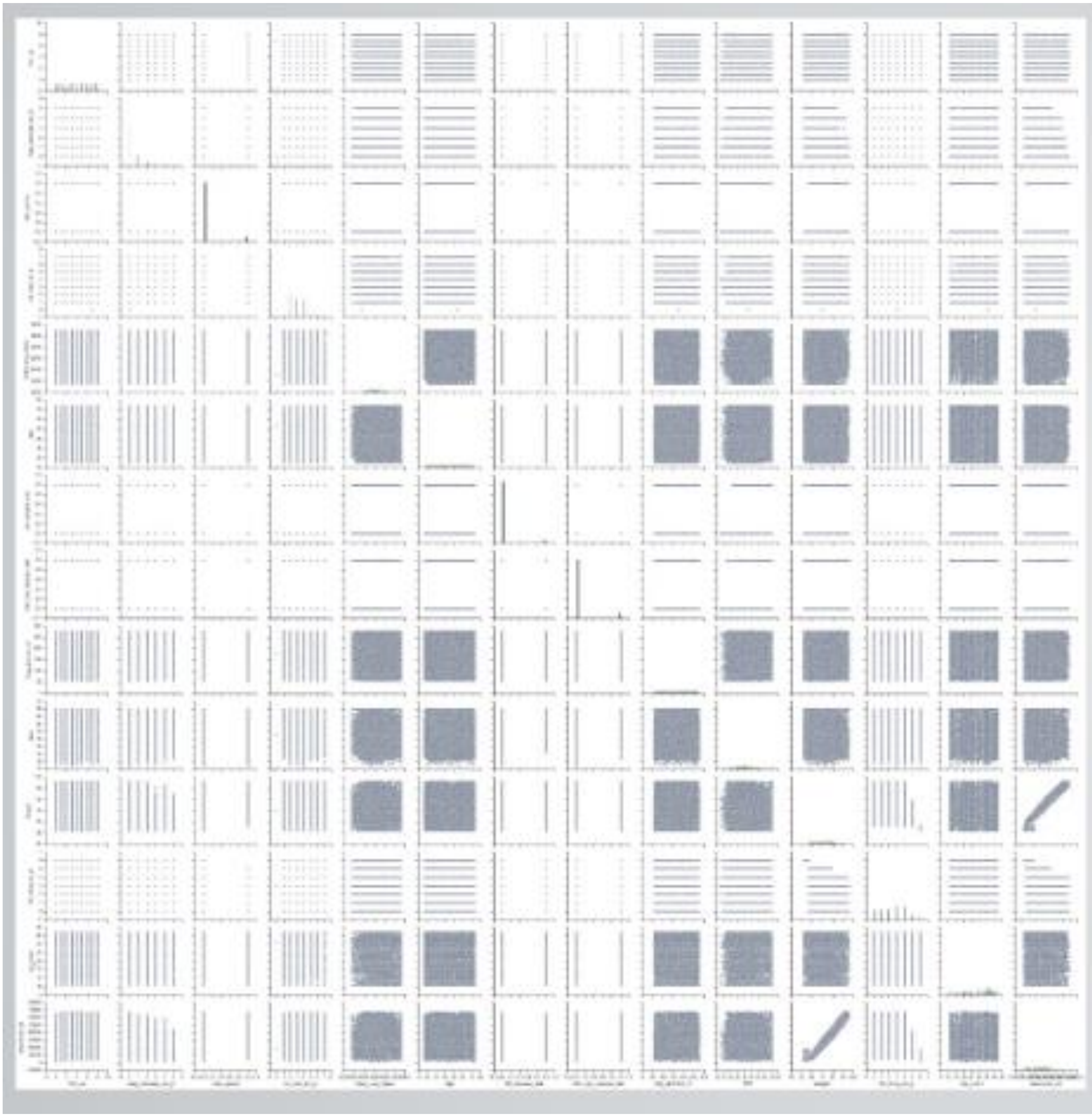
NULL VALUE TREATMENT:

CHECK FOR NULL VALUES:

BMI has **24010 non-null values**. Year_last_admtd has **13119 has non-null values**. Attributes BMI and Year_last_admtd has missing values. BMI has **990** missing values, which is **3.96%** of the entire data and Year_last_admtd has **11881** missing values, which accounts for **47.52%** of the entire data.

All the other columns do not have any missing values.

CORRELATION:



As it can be noticed, there is high Linear correlation between The feature weight and property. The rest of the features have very less correlation with the Property.

Thus, by looking at this graph, the correlation values and by using our knowledge in the field, several features are dropped during or before modelling to reduce dimensionality and thus reduce the load on the model.

ANALYSIS METHODOLOGY:

For analysis methodology we have used both regression and classification methodology.

In regression we have used **PCR, SVR, Ridge and LASSO regression.**

In Classification, we have used **Gradient Boosting classification.**

REGRESSION MODEL:

SVR:

We have used SVD for the accuracy and interpretability. In this process we have divided the data into train and test. Now the validation for the trained model against the test data. The accuracy in SVD is **95.1%.**

Model	RMSE Train	RMSE Test	Accuracy
SVR	3167.177	3199.231	95.1%

```
Call:
best.tune(METHOD = svm, train.x = prop_train_svr ~ descriptors_train_svr, ranges = list(epsilon = seq(0,
    1, 0.2), cost = 1:5))


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.04761905
    epsilon:  0.2


Number of Support Vectors:  7961
```

LASSO AND RIDGE REGRESSION:

Ridge regression is a tuning method which is used to analyze the data which suffers from overfitting. Ridge regression penalize the model by square of coefficient values.

LASSO regression penalizes the model by the absolute values of the coefficient values. The accuracy for ridge and lasso is 93.2% and 94.4%.

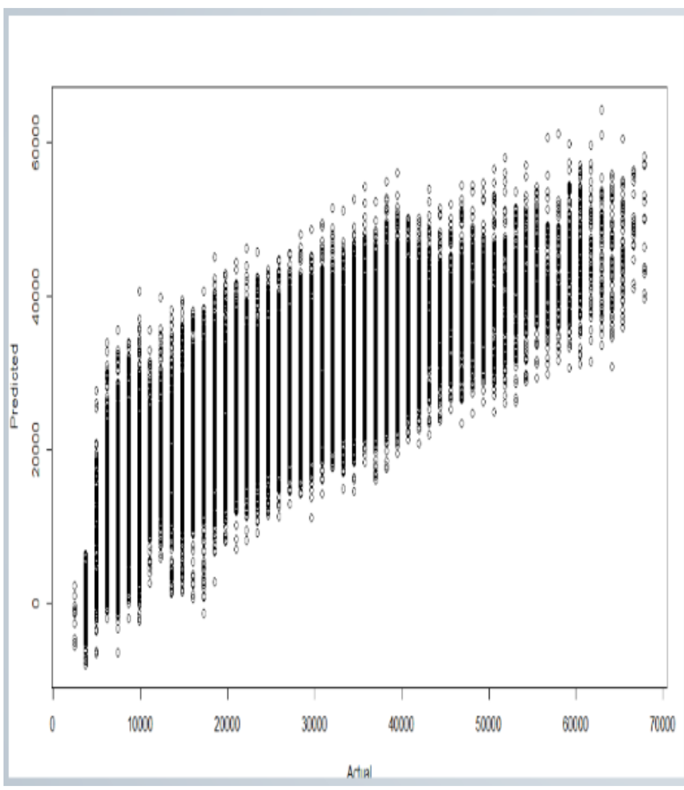
Model	RMSE Train	RMSE Test	Accuracy
Ridge regression	3718.853	3759.311	93.2%
Lasso regression	3362.604	3382.708	94.4%

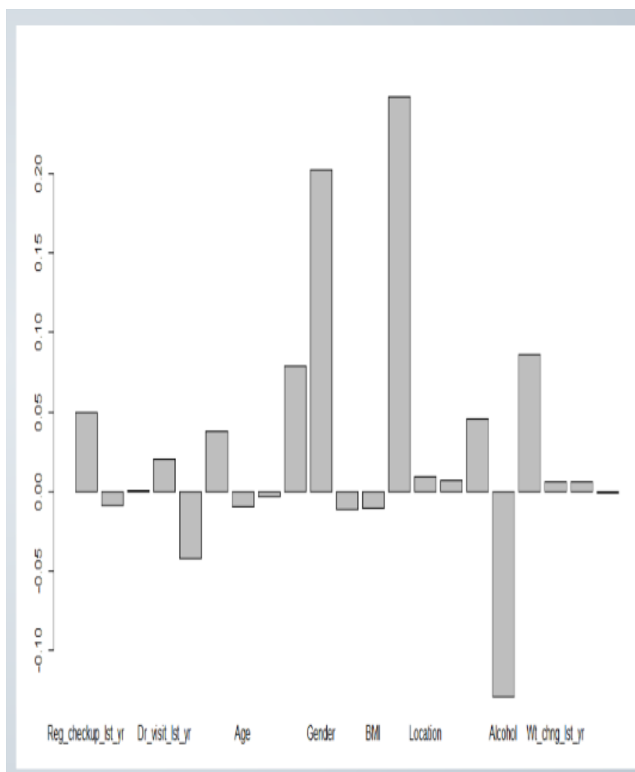
PCR:

Modeling the connection between a target variable and the predictor variables is the objective of principal component regression (PCR), a regression approach. When the data's greater variability and (presumably) its relevance to the objective variable are best represented by the lesser number of primary components. As a result, we only use a portion of the principal components for regression, rather than all the original characteristics. For PCA, more the PCR we will obtain more accuracy. In our case we have used 7 PCR. Beyond this the accuracy is not exceeding **65.35%** and remains the same. For 4 PC's, the average accuracy is **24.3%.**

Model	RMSE Train	RMSE Test	Accuracy
PCR(4PCs)	8827.812	8960.899	61.8%
PCR(5PCs)	8500.588	8445.047	64.7%
PCR(7PCs)	8422.509	8361.098	65.35%

PCA PLOT REPRESENTATION:



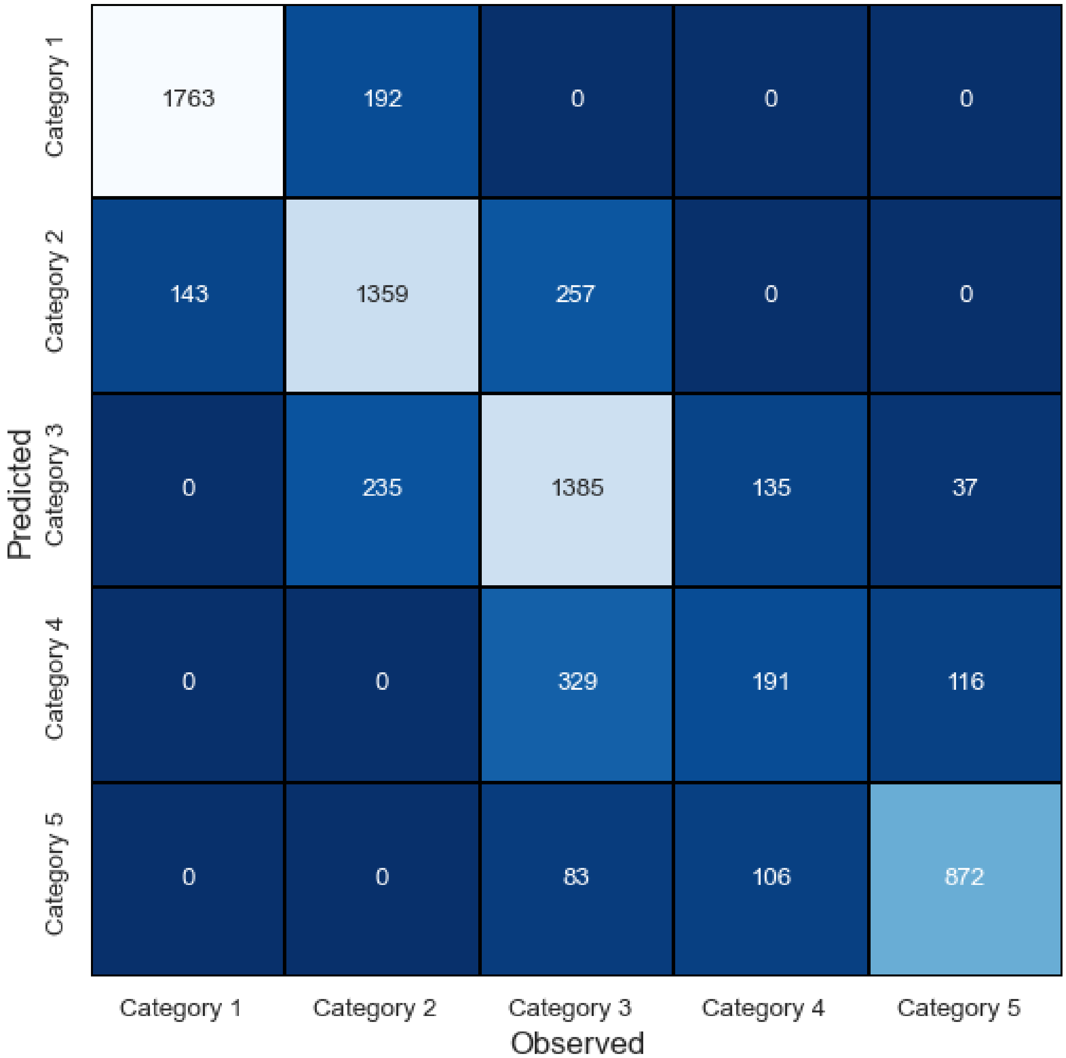


CLASSIFICATION MODEL:

XGBOOST:

XGBoost is a Gradient boosted Decision Tree that is used here for classification purpose. The goal is to predict the Insurance Premium range based on the features finalized. The values of **'Insurance_cst'** at binned at the [**0,.25,.5, .75, .85, 1]** percentiles. These percentiles are chosen upon analyzing the distribution of the feature 'Insurance_cst'. After fine-tuning the parameters (max_depth = 3, n-estimators=200), training accuracy of 81% and testing accuracy of 0.77% was achieved.

Confusion Matrix of Y_test vs Y_Predicted values:



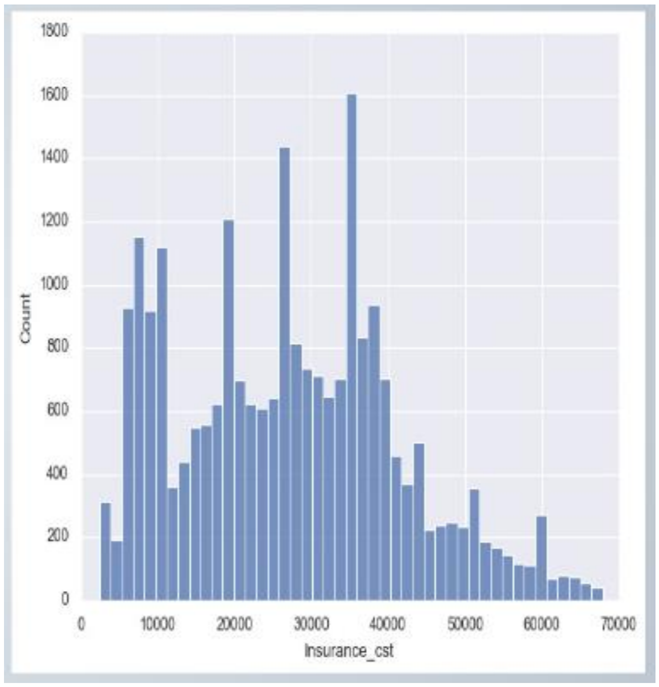
BINNING DATA

- 5 bins are created
- Bins are made based on amount of premium being paid.

CATEGORIES	NO OF RECORDS
0	6524
1	5843
2	6052
3	2098
4	3493

BOOSTING CLASSIFIER

Category	Precision	Recall	Support
1 (₹7500)	0.92	0.90	1955
2 (₹20000)	0.76	0.77	1759
3 (₹33500)	0.67	0.77	1792
4 (₹43000)	0.44	0.30	636
5 (₹57000)	0.85	0.82	1061



.

CONCLUSION:

Our data very nicely consists of data from almost all susceptible ages, and going forth one of the aims should be to create that kind of data distribution among all attributes. As an insurance company, along with helping our customers, our main aim is to create maximum profit while taking under consideration aspects that can potentially make us lose customers. So, the best customers for us are the ones who are fit but hardly ever require treatment, but even better customers are the ones that are over-weight but hardly claim insurance. We can conclude that for Regression, **SVR model** can be used as it has the best performance. The **Classification model** has a decent accuracy, but the performance can be improved using Over-sampling.

REFERENCE:

1.K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar and R. K. Bhatia, "Health Insurance Cost Prediction using Machine Learning," 2022 3rd International Conference for Emerging Technology (INCET), 2022, pp. 1-5, doi: 10.1109/INCET54531.2022.9824201.

2.M. A. Morid, K. Kawamoto, T. Ault, J. Dorius and S. Abdelrahman, "Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation", *AMIA Annual Symposium Proceedings*, vol. 2017, pp. 1312, 2017.Show in Context Google Scholar

3.Philipp Drewe-Boss, Dirk Enders, Jochen Walker and Uwe Ohler, "Deep learning for prediction of population health costs", *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1-10, 2022. Show in Context CrossRef Google Scholar.

4.C. A. Powers, C. M. Meyer, M. C. Roebuck and B. Vaziri, "Predictive modeling of total healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques", *Med. Care*, vol. 43, pp. 1065-1072, 2005.Show in Context CrossRef Google Scholar.

5.MC Politi, E Shacham, AR Barker, N George, N Mir, S Philpott et al., "A Comparison Between Subjective and ObjectiveMethods of Predicting Health Care Expenses to Support Consumers'". Show in Context Google Scholar.

6.*MDM Policy & Practice.*, vol. 3, no. 1, pp. 238146831878109, 2018.CrossRef Google Scholar .

7."Medical Cost Personal Datasets", [online] Available: https://www.kaggle.com/mirichoi0218/insurance.