Introduction to Machine Learning Graded Project

Week 3

Part I

Introduction

The objective of this whole exercise is to understand the concepts of probability, Bayes theorem, normal distribution, binomial distribution, poisson distribution and their applications.

Probability and Bayes Theorem

Q1. A consumer research survey sampled 200 men to find out whether they prefer to drink plain water or soft drink. 80 out of these 200 men prefer a soft drink. What is the probability that a randomly chosen man will prefer a soft drink?

Probability that a man prefers soft drink = number of men who prefer soft drink / total number of men in the sample.
Hence probability is 80/200 = 0.4

Q2. From a full deck of 52 cards, 1 card is drawn randomly. What is the probability that the card is either a spade or a king?

Total cards = 52
spades = 13
kings = 4
spade which is king = 1

probability = probability(king) + probability(spade) - probability(spade which is king)
= 4/52 + 13/52 — 1/52 = 0.308

Q3. A drilling company has estimated a 40% chance of striking oil for their new well. A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests. Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful?

Probability of success = 0.4
Probability of failure = 0.6
Probability (detailed test / success) = 0.6
Probability (detailed test / failure) = 0.2

Formula for probability that the well will be successful
prob = Probability of success * Probability (detailed test / success) / {Probability of success * Probability (detailed test / success) + Probability of failure * Probability (detailed test / failure) }
= 0.4 * 0.6 / {0.4*0.6 + 0.6* 0.2} = 0.667

Normal Distribution

Q1. There are 1000 students in a class. The average Score and the Variance of the Score of the class is 240 and 400 respectively. The Scores of the students are normally

distributed. Rahul, a student in the class belongs to the 95th percentile in the class. What is the actual Score Rahul has got?

Mean = 240
variance = 400
standart deviation = 20
For the sample to be equal to 95th percentile of the hole sample the z value is 1.645
Using this formula, let's find x (the actual Score Rahul has got)
x - mean / standart deviation = z
x = z * standart deviation + mean
x = 1.645 * 20 + 240
x = 272.9

Q2. The mean score on a college placement exam is 500 with a standard deviation of 100. Ninety-five percent of the test takers score above what?

mean = 500
standart deviation = 100
It is said that 95% if test takers score should be above this value.
Hence for first 5% the z value is -1.645

Using this formula, let's find x
x = z * standart deviation + mean
x = -1.645 * 100 + 500
x = 335.51

Q3. A speed-data of some cars is given. The speeds are normally distributed with a mean of 70 km/hr and a standard deviation of 10 km/hr.

a) What is the probability that a car picked at random is travelling at more than 100 km/hr?

mean = 70
standart deviation = 10
x = 100
To calculate the probability that a car picked at random is travelling at more than 100 km/h, firstly we calculate the probability of 100 km/hr and then substract it from 1

z = (x - mean) / standart deviation
z = 3
prob (x= 100) = 0.99865
prob (x > 100) = 1 - prob (x= 100) = 1 - 0.99865 = 0.00135

b) What is the probability that the car speed is between 80 Km / hr and 100 Km / hr

mean = 70
standart deviation = 10
x = 80

z = (70 - 80) / 10
z = 1
prob(80 < x < 100) = prob(z=3) - prob(z=1) =0.998650 - 0.841344 = 0.157305

Binomial Distribution

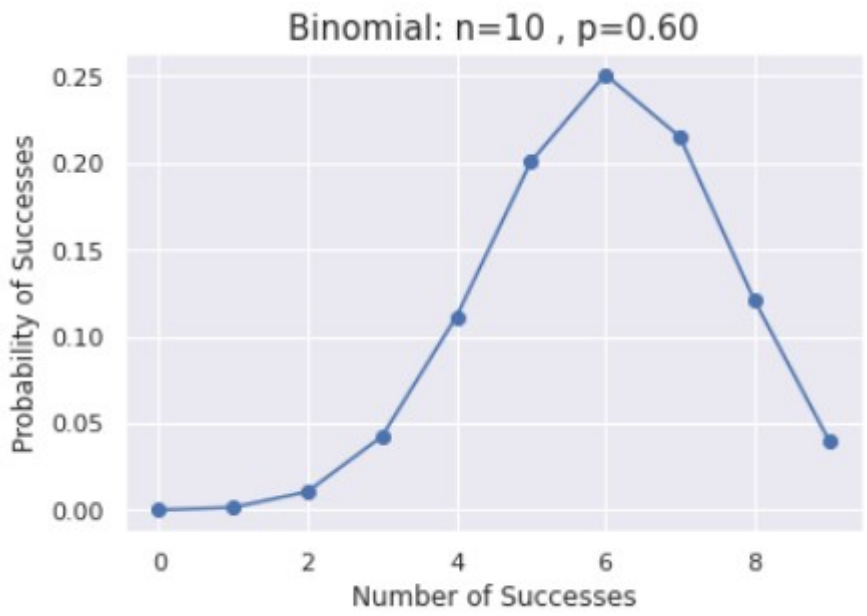
Q1. You flip a fair coin 10 times. What is the probability of getting 8 or more heads?
Hint - Use stats.binom.pmf() function for this.

Probability of getting 8 or more heads for flipping 10 coins is
prob of getting 8 heads + prob of getting 9 heads + prob of getting 10 heads

binomial(k=8, n=10, p=0.5) + binomial(k=9, n=10, p=0.5) + binomial(k=10, n=10, p=0.5)
= 0.0546

Q2. My Bank has a large Credit Card portfolio. Based on empirical data, they have found that 60% of the customers pay their bill on time. If a sample of 10 accounts is selected from the current database, construct the Probability Distribution of accounts paying on time.

binomial(k=np.arange(0,10), n=10, p=0.6)



Poisson Distribution

Q1. Assume a poisson distribution with lambda = 5.0. What is the probability that

a) X <= 1?
p(x <= 1) = poisson(x=0, lambda=5.0) + poisson(x=1, lambda=5.0) = 0.0404

b) X > 1?
p(x > 1) = 1 - p(x < = 1) = 1 - 0.0404 = 0.9595

Q2. The number of defects per month in a manufacturing plant is known to follow a Poisson distribution, with a mean of 2.5 defects a month.

a) What is the probability that in a given month, no defects occur?
probability(no defect) = poisson(x = 0, lambda = 2.5) = 0.082

b) That at least one defect occurs?
probability(at least one defect) = 1 - probability(no defect) = 0.917

Part 2

Insurance Claim Prediction

Context

A key challenge for the insurance industry is to charge each customer an appropriate premium for the risk they represent. The ability to predict a correct claim amount has a significant impact on insurer's management decisions and financial statements. Predicting the cost of claims in an insurance company is a real-life problem that needs to be solved in a more accurate and automated way. Several factors determine the cost of claims based on health factors like BMI, age, smoker, health conditions and others. Insurance companies apply numerous techniques for analyzing and predicting health insurance costs.

Data Description:

age : Age of the policyholder
sex: Gender of policyholder
bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight
children: Number of children of the policyholder
smoker: Indicates policy holder is a smoker or a non-smoker (non-smoker=0;smoker=1)
region: The region where the policy holder belongs to (northeast, northwest, southeast, southwest)
claim: Claim amount
bloodpressure: Blood pressure reading of policyholder
diabetes: Suffers from diabetes or not (non-diabetic=0; diabetic=1)
regular_ex: Regularly exercise or not (no-exercise=0; exercise=1)

1. Import the data and perform the following checks and write down your insights at every step.

a. Shape of the data
There are 1338 rows and 10 columns in the dataset

b. Data types of attributes

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   age            1338 non-null   int64
 1   sex            1338 non-null   object
 2   bmi            1338 non-null   float64
 3   children       1338 non-null   int64
 4   smoker         1338 non-null   int64
 5   region         1338 non-null   object
 6   bloodpressure  1338 non-null   int64
 7   diabetes       1338 non-null   int64
 8   regular_ex     1338 non-null   int64
 9   claim          1338 non-null   float64
dtypes: float64(2), int64(6), object(2)
memory usage: 104.7+ KB
```

The dataset consist float, int, and object data types.

c. 5-point summary of the relevant attributes

	age	bmi	children	smoker	bloodpressure	diabetes	regular_ex	claim
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.808281	1.094918	0.144245	68.977578	0.687593	0.319133	13270.422414
std	14.049960	6.282207	1.205493	0.351469	19.327770	0.463648	0.466315	12110.011240
min	18.000000	16.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1121.870000
25%	27.000000	26.300000	0.000000	0.000000	64.000000	0.000000	0.000000	4740.287500
50%	39.000000	30.500000	1.000000	0.000000	72.000000	1.000000	0.000000	9382.030000
75%	51.000000	34.800000	2.000000	0.000000	80.000000	1.000000	1.000000	16639.915000
max	64.000000	62.000000	5.000000	1.000000	122.000000	1.000000	1.000000	63770.430000

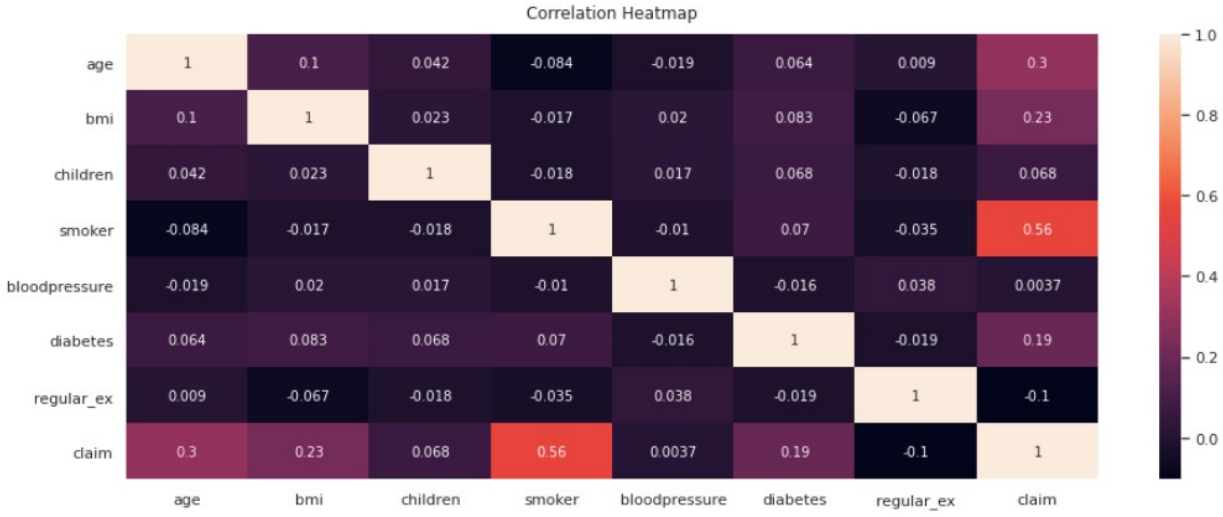
Max value of claim is too far from 75th percentile. This seems to have a lot of outliers there.

d. Missing values

```
age             0
sex             0
bmi             0
children        0
smoker          0
region          0
bloodpressure   0
diabetes        0
regular_ex      0
claim           0
dtype: int64
```

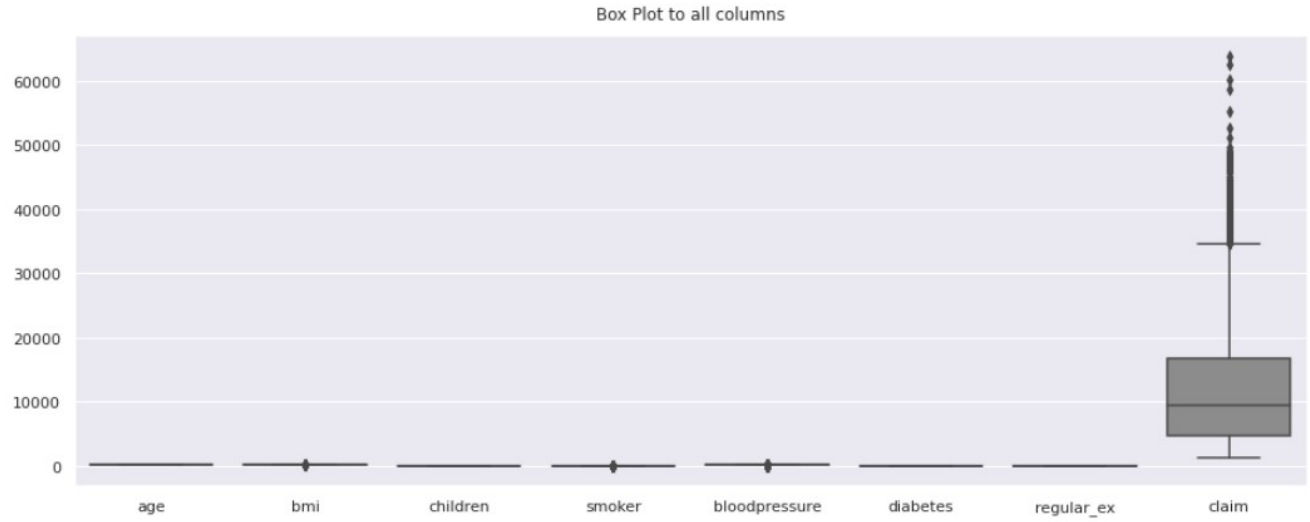
There are no missing values in the dataset

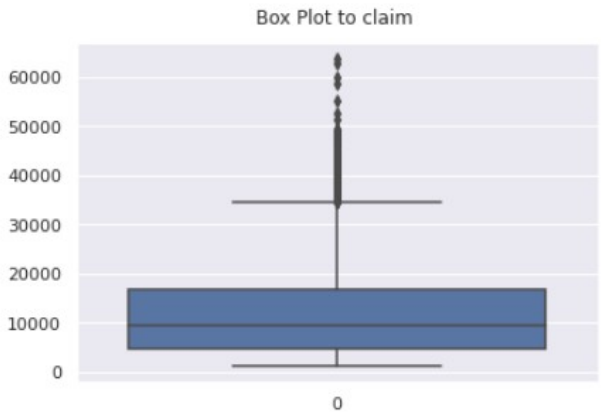
e. Correlation among the attributes

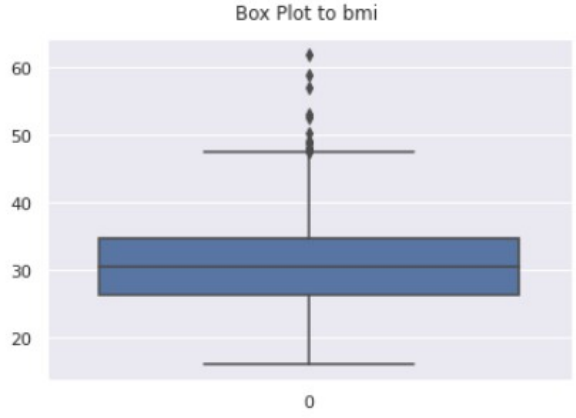


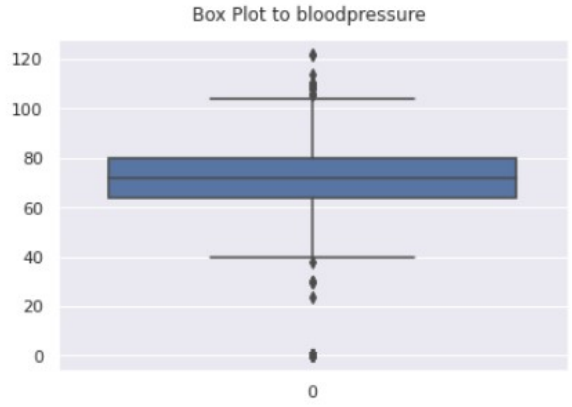
There is a 0.56 correlation smoker to claim colum. I suppose that smoker will have high weight in linear regression equation.

f. Outliers (display a boxplot)
From the below boxplot we can see that bmi, bloodpressure and claim attributes have outliers.

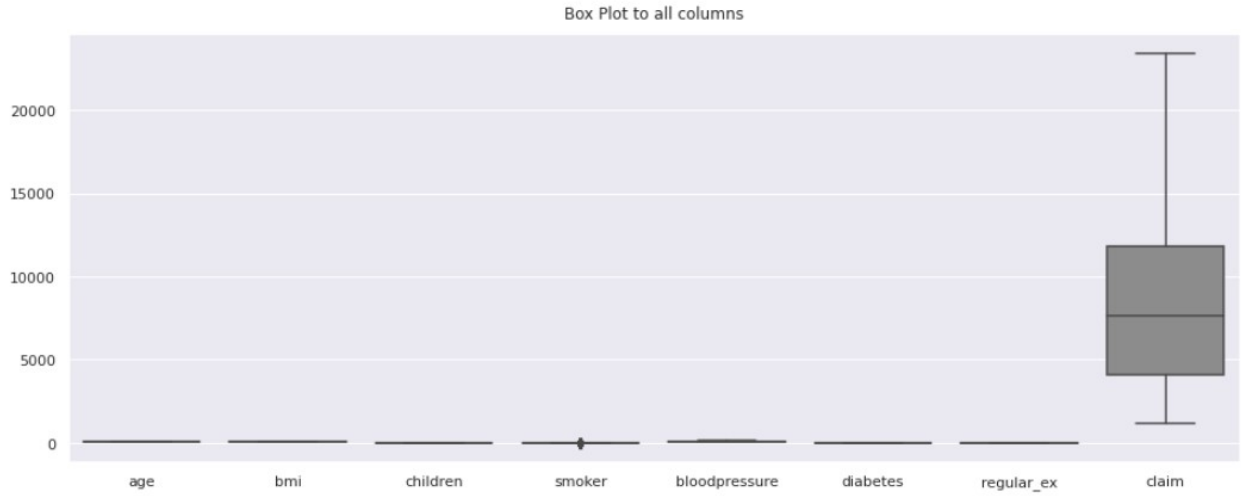








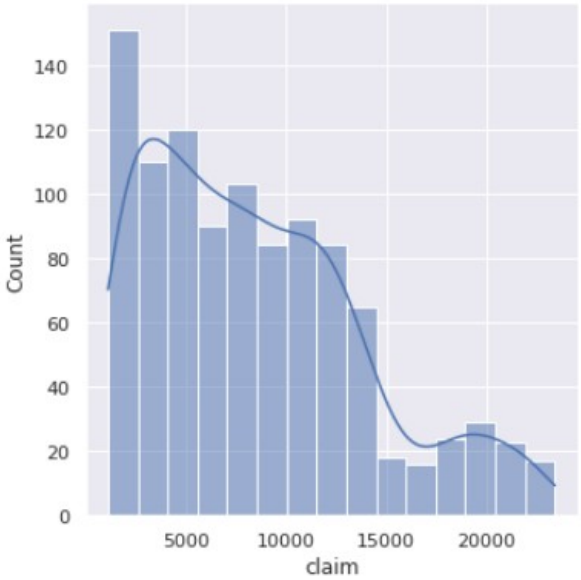
g. Remove outliers (using IQR)

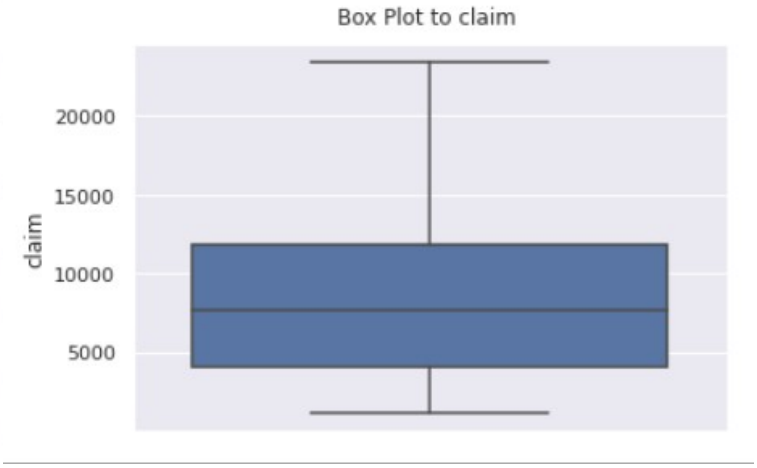


From the above figure we can see that all the outliers are removed.
Outlier from bmi, bloodpressure are dropped by calculating IQR and outliers values for the claim attribute are changed to lower and upper ranges using IQR.

h. Distribution of the target column(“claim”)

The above figure shows the distribution of claim attribute.



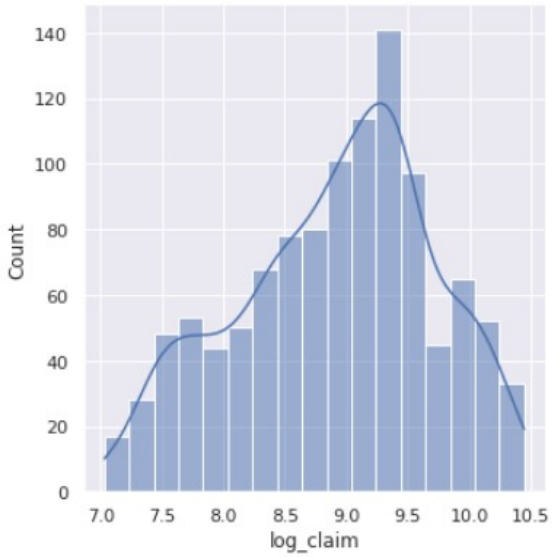


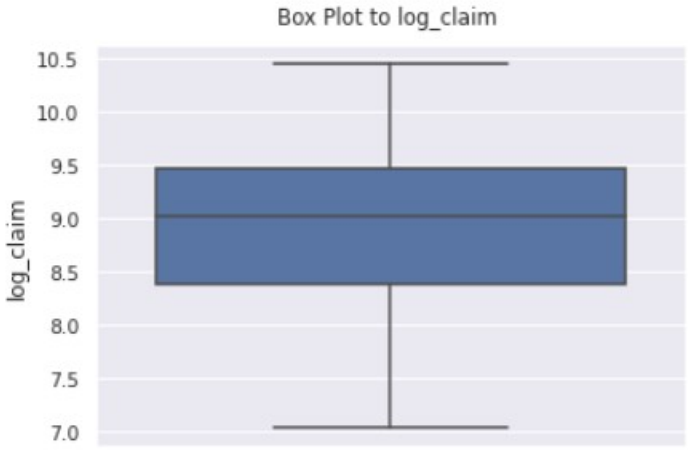
2. Transform the column “claim” using log transformation (hint: use np.log(‘column’) and append the transformed column to the dataframe under the column name “log_claim” - optionally you can check the effect of the transformation by plotting histogram of “claim” before and after transformation.

Kurtosis of claim is -0.13976466696270684
Skewness of claim is 0.7408945365961965

Kurtosis of log_claim is -0.6923468126647134
Skewness of log_claim is -0.45015410578357434

The skewness of the claim is reduced by doing log transform, It changed from right skewed to left skewed.





3. Encode the categorical variables. In case a column has more than 2 categories, use one-hot encoding.

From the above attributes, for sex, region attributes encoding is needed, because linear model cannot fit object data.

Encoding of sex attribute is done by mappping female category with zero and male category with 1.
Encoding of region attribute is done by creating dummies i.e., one-hot encoding.

	age	sex	bmi	children	smoker	bloodpressure	diabetes	regular_ex	claim	northeast	northwest	southeast	southwest
0	48	1	28.0	1	1	66	0	0	23568.27	0	0	0	1
1	39	1	34.1	2	0	70	0	1	23563.02	0	0	1	0
2	47	0	26.1	1	1	80	1	0	23401.31	1	0	0	0
3	49	1	25.6	2	1	62	0	0	23306.55	0	0	0	1
4	53	0	22.9	1	1	82	0	0	23244.79	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1023	18	1	34.4	0	0	55	1	0	1137.47	0	0	1	0
1024	18	1	34.1	0	0	90	1	0	1137.01	0	0	1	0
1025	18	1	33.7	0	0	80	1	1	1136.40	0	0	1	0
1026	18	1	33.3	0	0	80	1	1	1135.94	0	0	1	0
1027	18	1	30.1	0	0	64	0	0	1131.51	0	0	1	0

4. Separate out the dependent variable(“claim”) from the independent variables(exclude claim and log_claim from the rest of the variables).

By dropping claim and log_claim columns the remaining data is taken as independent variables taking it as X.
the claim column is taken as Y i.e., dependent variable.
log_claim column is taken as Z i.e., dependent variable.

5. Split the data into testing and training sets (X_train, y_train, X_test, y_test).

Splitting of X and y data is done using train_test_split imported from sklearn.model_selection

X_train : 746 rows , 12 columns
X_test : 368 rows, 12 columns

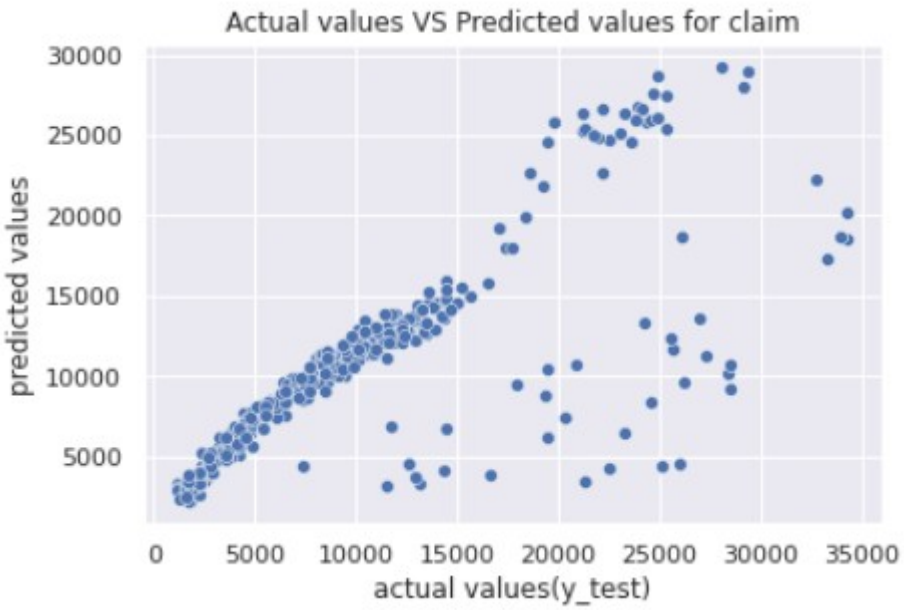
y_train : 746 rows
y_test : 368 rows

6. Train a linear regression model using the training data and print the r_squared value of the prediction on the test data.

Training data is fit into linear regression model which is LinearRegression() imported from sklearn.linear_model

On test data R_squared is 0.5891548396602004

7. Plot a scatter plot between the actual values and the predicted values for the test set (because plain numbers might not give the entire picture).

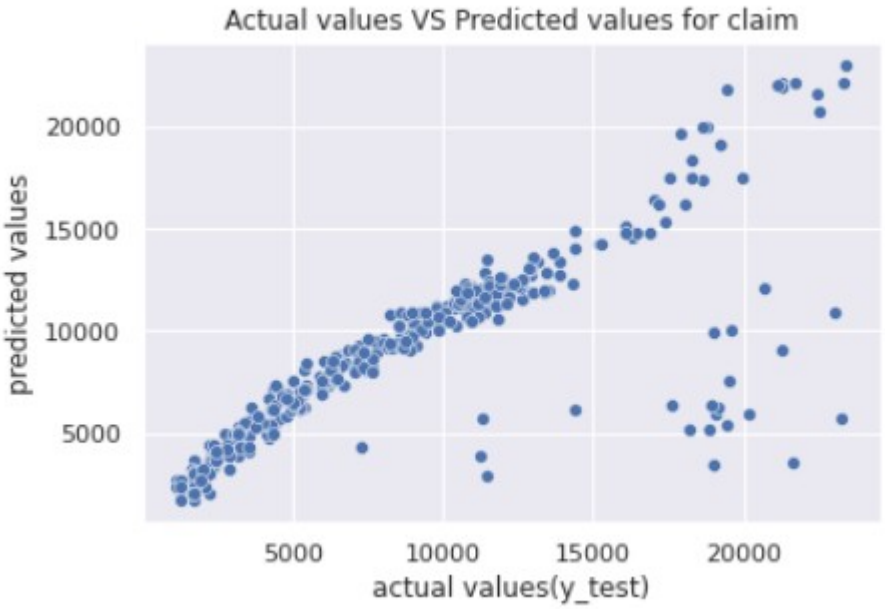


The above figure gives the scatter plot between predicted data and actual values of test data.

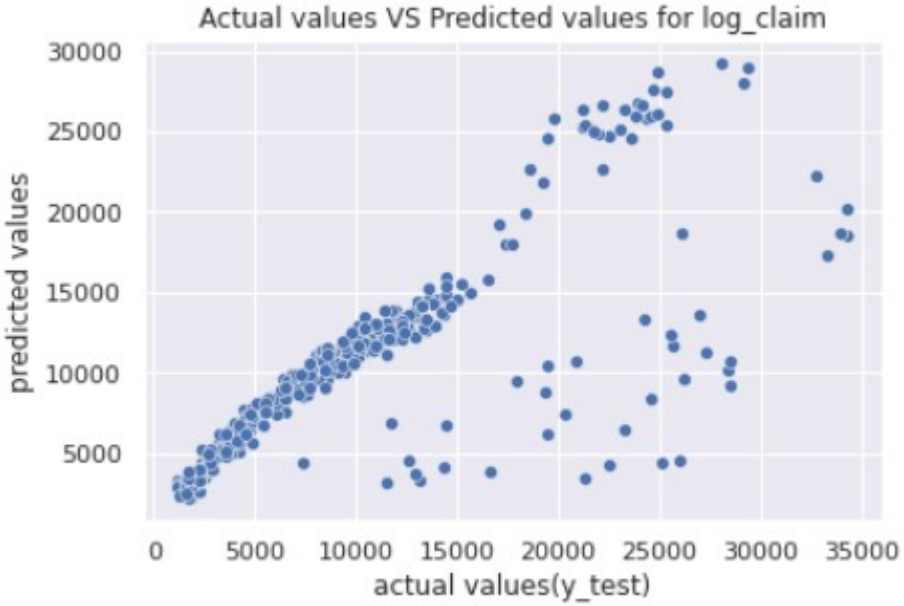
8. Comment on the performance of the model.

Since the R squared value is 0.59 for test data which is low we can say that the linear fit done for this data is not a really good fit. and also from the scatter plot we see that there is much difference between predicted data and actual values.
At the beggining we can notice a pattern of a slightly curved line but in the middle there are few scatter clouds. For further analysis I would like to try polynomial regression model.
In my polynomial regression model with degree 2, I got worse results than I expected to get. R squared value is 0.58 for test data

For an experiment I dropped outliers few times and I get better results. With more manipulation with data R squared value became a little higher 0.63 for test data.

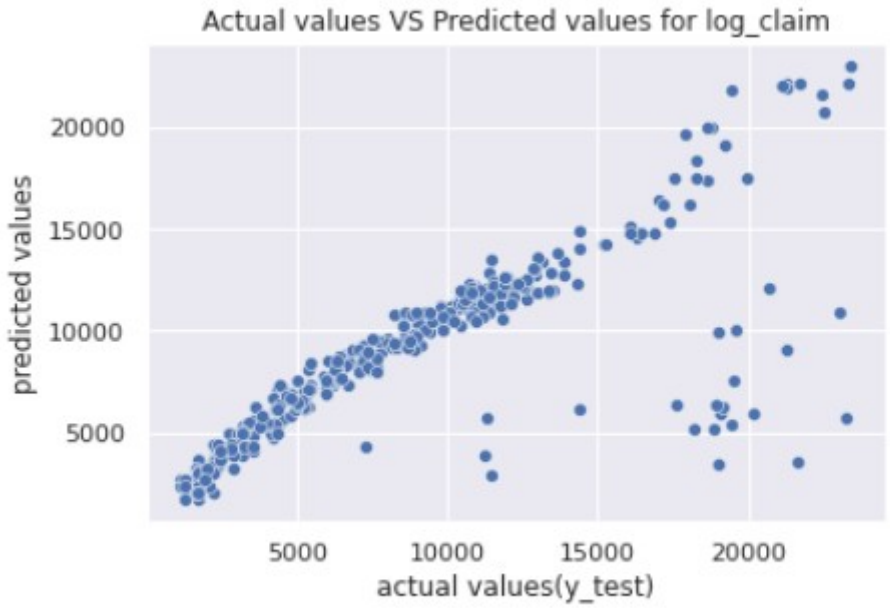


9. Repeat steps 4, 5, 6,7 and 8 except, this time use “log_claim” as your dependent variable (note: “claim” cannot be among the predictors).



R squared value is 0.65 for test data with log_claim

For an experiment I dropped outliers few times and I get better results. With more manipulation with data R squared value became a little higher 0.71 for test data.



10. Compare the performance of the models trained using the skewed dependent variable as it is and log transformed variable - write your comments and conclude the project.

Comparing the performance of linear regression model for given dependent variable and log transformed dependent variable we can see that R squared value is increased by decreasing the skewness by doing log transformation, that is the log transformed variable regression is a good fit compared to the normalvariable regression.

For better results we can delete outliers several times, after this manipulation linear reggression with logged dependent values is showing good result of 0.71 for test data.

Hence we can say that decrease in skewness causes the over fit regression hence causing high change in the R squared value between the train data set and test data set that is model validation for test data is decreasing. The same can be observed from the scatter plots.