PREDICTION OF INSURANCE CLAIMS

USING HEALTH ANALYSIS



PROJECT DONE BY:

UVA PRASANTH . P

AUSTRO FLOVIN .V

JEFFRY DANIEL JEBAKUMAR

INDEX

1. Title of the project
   1. Introduction
   2. Problem statement
   3. Abstract & objective

2. Data collection

3. Methodology used

3.1 Data distribution analysis

3.2 Average cost analysis

3.3 Training the data

3.4 Error calculation

3.5 Feature importance value

3.6 Prediction of new data

4. Finding and suggestions

5. Conclusion & Reference

1. Title of the Project : Prediction of insurance claims using health analysis
   1. Introduction

A health insurance policy is a contract between an insurance provider (e.g. an insurance company or a government) and an individual or his/her sponsor (e.g. an employer or a community organization). The contract can be renewable (e.g. annually, monthly) or lifelong in the case of private insurance, or be mandatory for all citizens in the case of national plans. The type and amount of health care costs that will be covered by the health insurance provider are specified in writing, in a member contract or "Evidence of Coverage" booklet for private insurance, or in a national health policy for public insurance.(US specific) Provided by an employer-sponsored self-funded ERISA plan. The company generally advertises that they have one of the big insurance companies. However, in an ERISA case, that insurance company "doesn't engage in the act of insurance", they just administer it. Therefore, ERISA plans are not subject to state laws. ERISA plans are governed by federal law under the jurisdiction of the US Department of Labor (USDOL). The specific benefits or coverage details are found in the Summary Plan Description (SPD). An appeal must go through the insurance company, then to the Employer's Plan Fiduciary. If still required, the Fiduciary's decision can be brought to the USDOL to review for ERISA compliance, and then file a lawsuit in federal court.

The Indian constitution makes the provision of health care in India the responsibility of the state governments, rather than the central federal government. It makes every state responsible for "raising the level of nutrition and the standard of living of its people and the improvement of public health as among its primary duties". The recent four main updates in 2017 mentions the need to focus on the growing burden of non-communicable diseases, on the emergence of the robust healthcare industry, on growing incidences of unsustainable expenditure due to health care costs and on rising economic growth enabling enhanced fiscal capacity.[[3]](https://en.wikipedia.org/wiki/Healthcare_in_India#cite_note-:0-3) In practice however, the private healthcare sector is responsible for the majority of healthcare in India, and most healthcare expenses are paid directly out of pocket by patients and their families, rather than through health insurance. Government health policy has thus far largely encouraged private sector expansion in conjunction with well-designed but limited public health programmes.According to the World bank, the total expenditure on health care as a proportion of GDP in 2015 was 3.89%. Out of 3.89%, the governmental health expenditure as a proportion of GDP is just 1%, and the out-of-pocket expenditure as a proportion of the current health expenditure was 65.06% in 2015.

1.2.Problem statement

To Predict the cost that a person can claim for an insurance company when he/she met with an accident or falls ill. The prediction estimates the cost by considering the attributes like age,sex,BMI,smoker,children and region.

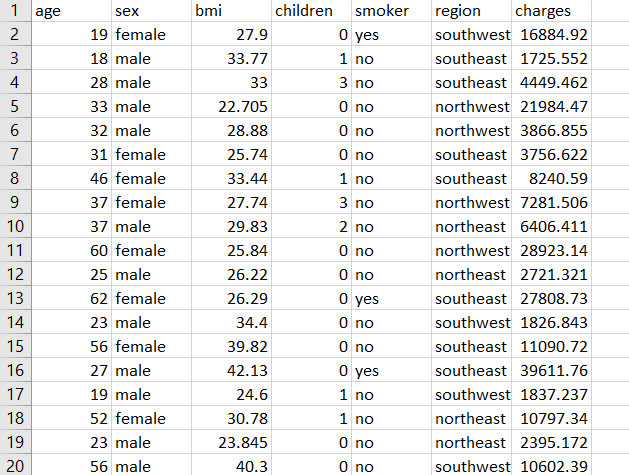
1.3.Abstract & objective

A study revealed that participants overall received only 54.9% of recommended care for 30 different acute and chronic conditions in US. This mandates a need for improvement in the management of chronic illnesses.The Patient-Centered Medical Home was first introduced in 1967 by the American Academy of Pediatrics and has evolved since that time. In 2007 the American Academy of Family Physicians, American Academy of Pediatrics, American College of Physicians, and American Osteopathic Association joined together to issue a statement on the Joint Principles of the Patient-Centered Medical Home (PCMH). This statement focused on seven main principles which were aimed to improve the landscape of disease management in the primary care setting. We constructed predictive models to predict the cost that a person can claim from an insurance company when he/she met an accident or falls ill. The main outcome of the project is “basing on the effect of attributes on the health of an individual, we predict how much can he claim from the insurance policy”.

2.Data collection

The dataset has various attributes on which the target variable i.e., the cost depends. The various independent attributes in the dataset are age, sex, body mass index, children, smoker and region. The insurance depends on these independent attributes and based on these attributes the cost is predicted.

The dataset that is used for our project is as follows:



The dataset consists of 7 attributes as mentioned before. The age ranges from 18 to 65. The region is categorized into 4 (northeast,northwest,southeast,southwest)

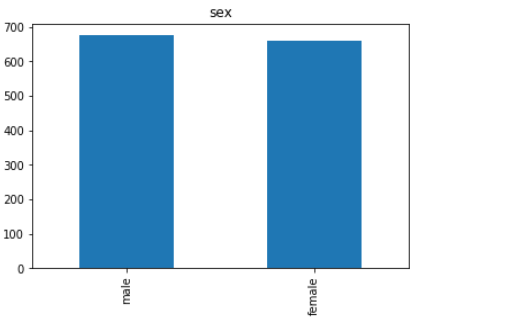
The dataset holds upto 1338 rows consisting of different samples.

3.Methodology used

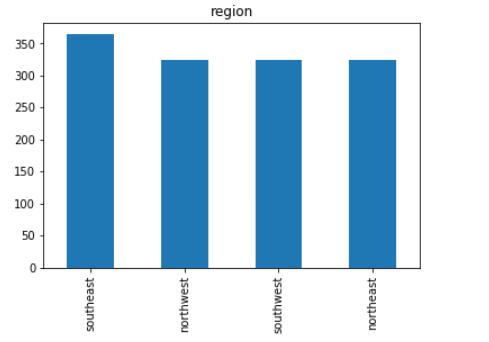
3.1.Data distribution analysis

To analyse how the data in the data set has been distributed , a bar graph has been plotted .

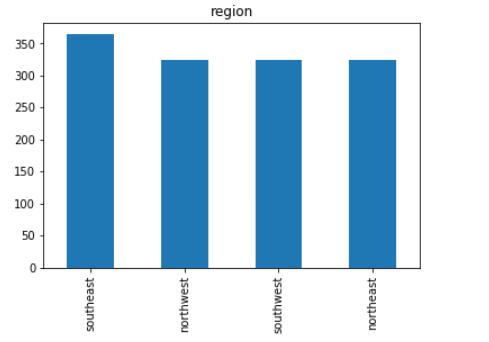
Data distribution analysis for sex.



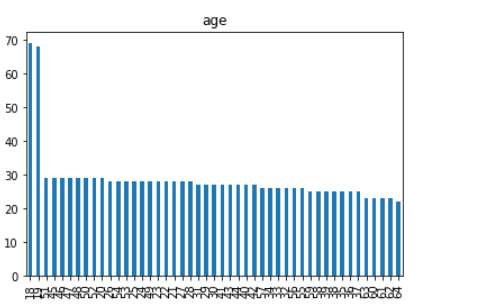
Analysis for region



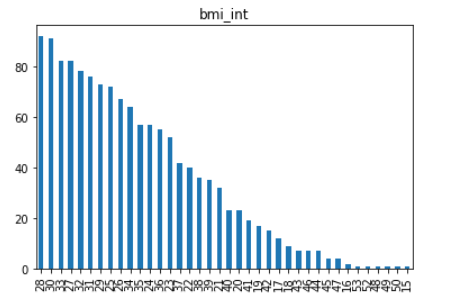
Analysis for smoker



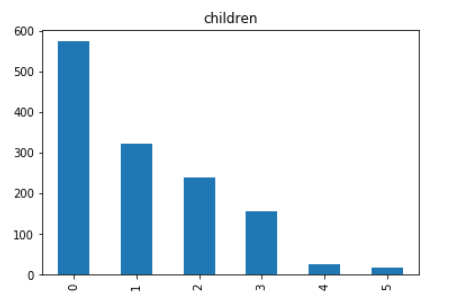
Analysis for age



Analysis for BMI



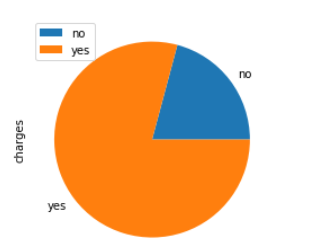
Analysis for children



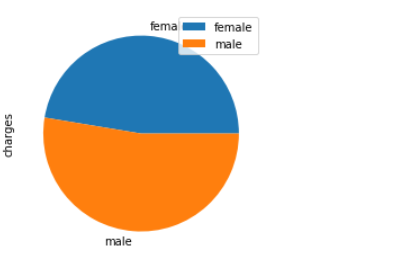
3.2.Average cost analysis

To predict and visualize the average cost for each attribute ,pie chart is used .

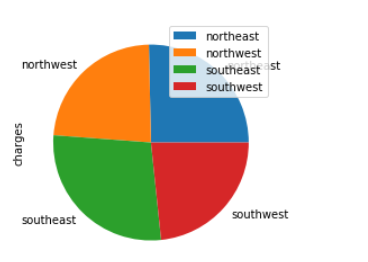
Cost analysis for smoker:



Cost analysis for sex:



Cost analysis for region:



3.3.Training the data:

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size=0.2)

regressor = ExtraTreesRegressor(n\_estimators = 200)

regressor.fit(X\_train,y\_train)

Algorithms learn from data. They find relationships, develop understanding, make decisions, and evaluate their confidence from the training data they’re given. And the better the training data is, the better the model performs.

In fact, the quality and quantity of your training data has as much to do with the success of your data project as the algorithms themselves.

Now, even if you’ve stored a vast amount of well-structured data, it might not be labeled in a way that actually works for training your model. For example, autonomous vehicles don’t just need pictures of the road, they need labeled images where each car, pedestrian, street sign and more are annotated; sentiment analysis projects require labels that help an algorithm understand when someone’s using slang or sarcasm; chatbots need entity extraction and careful syntactic analysis, not just raw language.

3.4.Error calculation

The errors in the linear regression is of two types. They are

1.Mean absolute error(MAE)

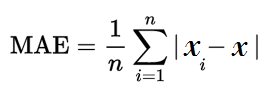
2.Root mean squared error(RMSE)

Mean absolute error:

In statistics **mean absolute error (MAE)** is a measure of difference between two continuous variable.The MAE is calculated between y\_trained value and y\_predicted value.

print("Train MAE: ", sklearn.metrics.mean\_absolute\_error(y\_train, y\_train\_pred))

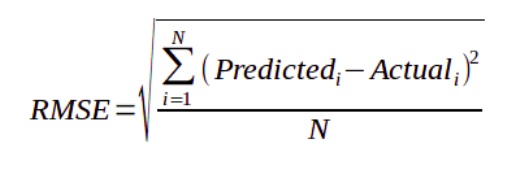
print("Test MAE: ", sklearn.metrics.mean\_absolute\_error(y\_test, y\_test\_pred))



Root mean squared error:

**Root Mean Square Error** (**RMSE**) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; **RMSE** is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

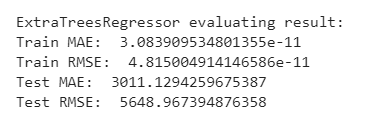
The root mean square error is calculated between the y\_trained value and y\_predicted value and also for y\_tested value and y\_predicted value.



print("Train RMSE: ", np.sqrt(sklearn.metrics.mean\_squared\_error(y\_train, y\_train\_pred)))

print("Test RMSE: ", np.sqrt(sklearn.metrics.mean\_squared\_error(y\_test, y\_test\_pred)))

Evaluated output:



3.5.Feature importance value

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

Having too many irrelevant features in your data can decrease the accuracy of the models.

Three benefits of performing feature selection before modeling your data are:

* **Reduces Overfitting**: Less redundant data means less opportunity to make decisions based on noise.
* **Improves Accuracy**: Less misleading data means modeling accuracy improves.
* **Reduces Training** Time: Less data means that algorithms train faster.

We measure the importance of a feature by calculating the increase in the model’s prediction error after permuting the feature. A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is “unimportant” if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction.

#Program for predicting the feature

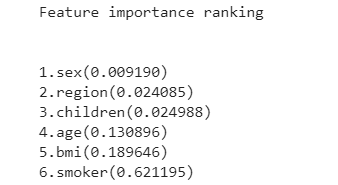
plt.figure()

plt.title("Feature importances")

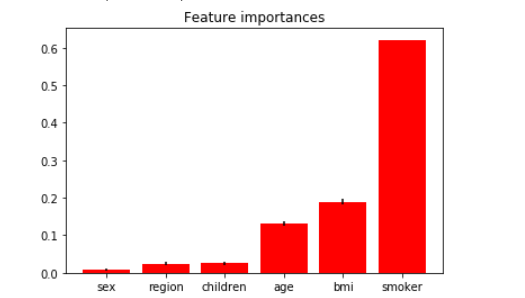
plt.pie(importance\_list, importances[indices], color="r", yerr=std[indices], align="center")

plt.show()

#Results of important feature prediction



#Bar graph for the prediction of important feature



3.6.Prediction of new data

With the given dataset and the trained ,tested value ,the prediction of cost for an individual has been calculated .

austro = ['male','yes','southeast',25,30.5,2]

print('austro - ',str(austro))

austro[0] = le\_sex.transform([austro[0]])[0]

austro[1] = le\_smoker.transform([austro[1]])[0]

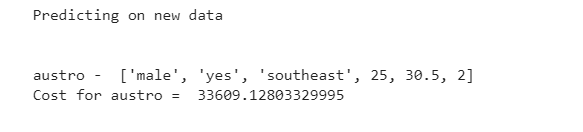
austro[2] = le\_region.transform([austro[2]])[0]

X = sc.transform([austro])

cost\_for\_austro = regressor.predict(X)[0]

print('Cost for austro = ',cost\_for\_austro,'\n\n')

Results of the new predicted data



4.Findings and suggestions

• Age plays a major role in the cost of a premium for health insurance. Younger people get less insurance, as the age increases their insurance also gets increased.

• Basing on the gender females get less insurance than males.

• Basing on the BMI, person with normal weight gets highest insurance, then the people with underweight, then the people with overweight, and the least insurance is for the people with obesity.

• When a person has no children or more than 3 children the insurance claiming will be less.

• Basing in the habit, if a person smokes he will get more insurance than the non smokers.

• Basing on the climatic conditions, if it is adverse the health may deteriorate so that the insurance claimed by the person will be more and vice versa.

Suggestions to Insurance Companies

• Creating more awareness regarding health insurance.

• Strong underwriting and claims management.

• Review of Mediclaim to cover ‘existing illness’, if possible with a higher premium.

• Introduction of new products for different market segments.

• Offer products for specific treatments to profitable segments.

Suggestions to Health insurance customers

• Taking health policy at very young age and covering all members of the family.

• Customers should be fully aware of the various health coverages available.

• Customers should know about the various health insurance schemes and companies providing these schemes.

• The attitude of customers should be always towards the preventive health care.

• Customers should take decisions relating to the features of the policy, sum assured, premium paid, persons covered, after careful analysis.

• They must be aware of the conditions and exclusions in the policy.

Program

#importing the library

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

from sklearn.ensemble import ExtraTreesRegressor

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn.model\_selection import train\_test\_split

import numpy as np

import sklearn.metrics

#load data

df = pd.read\_csv('insurance.csv')

df = df.dropna()

#general information

df.describe()

df.corr()

df['bmi\_int'] = df['bmi'].apply(lambda x: int(x))

variables = ['sex','smoker','region','age','bmi\_int','children']

# data distribution analysys

print('Data distribution analysys')

for v in variables:

df['bmi\_int'] = df['bmi'].apply(lambda x: int(x))

df[v].value\_counts().plot(kind = 'bar')

plt.title(v)

plt.show()

#average cost analysys

print('Mean cost analysys:')

for v in variables:

group\_df = df.groupby(pd.Grouper(key=v)).mean()

group\_df = group\_df.sort\_index()

group\_df.plot(y = ['charges'],kind = 'pie')

plt.show()

#training

print('Model training and evaluating\n\n')

#Transform categorical data ,text data into numbers

le\_sex = LabelEncoder()

le\_smoker = LabelEncoder()

le\_region = LabelEncoder()

df['sex'] = le\_sex.fit\_transform(df['sex'])

df['smoker'] = le\_smoker.fit\_transform(df['smoker'])

df['region'] = le\_region.fit\_transform(df['region'])

variables = ['sex','smoker','region','age','bmi','children']

X = df[variables]

sc = StandardScaler()

X = sc.fit\_transform(X)

Y = df['charges']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size=0.2)

#train model

regressor = ExtraTreesRegressor(n\_estimators = 200)

regressor.fit(X\_train,y\_train)

#prediction and evaluation

y\_train\_pred = regressor.predict(X\_train)

y\_test\_pred = regressor.predict(X\_test)

print('ExtraTreesRegressor evaluating result:')

print("Train MAE: ", sklearn.metrics.mean\_absolute\_error(y\_train, y\_train\_pred))

print("Train RMSE: ", np.sqrt(sklearn.metrics.mean\_squared\_error(y\_train, y\_train\_pred)))

print("Test MAE: ", sklearn.metrics.mean\_absolute\_error(y\_test, y\_test\_pred))

print("Test RMSE: ", np.sqrt(sklearn.metrics.mean\_squared\_error(y\_test, y\_test\_pred)))

print('Feature importance ranking\n\n')

importances = regressor.feature\_importances\_

std = np.std([tree.feature\_importances\_ for tree in regressor.estimators\_],axis=0)

indices = np.argsort(importances)[::1]

importance\_list = []

for f in range(X.shape[1]):

variable = variables[indices[f]]

importance\_list.append(variable)

print("%d.%s(%f)" % (f + 1, variable, importances[indices[f]]))

# Plot the feature importances of the forest

plt.figure()

plt.title("Feature importances")

plt.bar(importance\_list, importances[indices],

color="r", yerr=std[indices], align="center")

plt.show()

print('Predicting on new data\n\n')

austro = ['male','yes','southeast',25,30.5,2]

print('austro - ',str(austro))

austro[0] = le\_sex.transform([austro[0]])[0]

austro[1] = le\_smoker.transform([austro[1]])[0]

austro[2] = le\_region.transform([austro[2]])[0]

X = sc.transform([austro])

cost\_for\_austro = regressor.predict(X)[0]

print('Cost for austro = ',cost\_for\_austro,'\n\n')

dennis = ['female','no','southeast',45,19,0]

print('Dennis - ',str(dennis))

dennis[0] = le\_sex.transform([dennis[0]])[0]

dennis[1] = le\_smoker.transform([dennis[1]])[0]

dennis[2] = le\_region.transform([dennis[2]])[0]

X = sc. transform([dennis])

cost\_for\_dennis = regressor.predict(X)[0]

print('Cost for Dennis = ',cost\_for\_dennis)

5.Conclusion

“Health coverage to all” should be the motto of the health insurance sector. There should be easy access to healthcare facilities and cost control measures should be in place. Health insurance is going to develop more in the current liberal economic scenario. But, a completely unregulated or very less regulated health insurance sector may concentrate only on those who have the ability to pay for the insurance cover. So, the challenge is in helping the benefits percolate to the economically weaker sections of the population. Transparent and accountable government and non-government participation should be encouraged. Developing and marketing social health insurance schemes through cooperatives and rural association would go a long way in benefiting the vast unorganized employment sectors currently neglected under the existing schemes. Also a thorough revamp of schemes like ESIS and CGHS is necessary for them to be more purposeful and efficient. If the government, service provider, health care industry and the health insurance customers can incorporate all these suggestions given in the study, then the concept of health insurance will reach new heights in the near future and Mother India will be definitely, the most healthiest nation in the world.

Reference

1. The Centers for Medicare & Medicaid Services (CMS) DoHaHS, United States. National Health Expenditure Data 2016. Available from : https://www.cms.gov/Research-StatisticsData-and-Systems/Statistics- Trends-and-Reports/NationalHealthExpendData/index.html.

2. Duncan I, Loginov M, Ludkovski M. Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. North American Actuarial Journal. 2016;20(1):65–87. [Google Scholar]

3. Lahiri C, Agarwal N. Predicting healthcare expenditure increase for an individual from medicare data. Proceedings of the ACM SIGKDD Workshop on Health Informatics. 2014 [Google Scholar]

4. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. Bmj. 2005;330(7482):68. [PMC free article] [PubMed] [Google Scholar]