Soft Computing Techniques (INT246)
Project Report

Insurance Cost Prediction

Submitted to: **Dr Aditya Khamparia**

Submitted by:

Sr. No.	Registration No	Name of Student	Roll No
1	11802671	Sanjay Das	A12
2	11802710	Abhisek Rautaray	A13

Acknowledgement

I would like to express my gratitude towards my University as well as teacher for providing me the opportunity to do this project. As a result, I came to know about so many new things. So, I am very much thankful to them.

Moreover I would like to thank my friends who helped me a lot whenever I got stuck in some problem related to my course. I am really thankful to have such a good support of them as they always have my back whenever I need.

Also, I would like to mention the support system and consideration of my parents who have always been there in my life to make me choose right thing and oppose the wrong. Without them I could never had learned and became a person who I am now.

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

Sanjay Das
11802671

Abhisek Rautaray
11802710

Library Used

In this project standard libraries for database analysis and model creation are used. The following are the libraries used in this project.

1. tkinter: It's a standard GUI library of python. Python when combined with tkinter provides fast and easy way to create GUI. It provides powerful object-oriented tool for creating GUI

It provides various widgets to create GUI some of the prominent ones being:

- Button
- Canvas
- Label
- Entry
- Check Button
- List box
- Message
- Text
- Messagebox

Some of these were used in this project to create our GUI namely messagebox, button, label, Option Menu, text and title. Using tkinter we were able to create an interactive GUI for our model.

1. Numpy: Numpy is core library of scientific computing in python. It provides powerful tools to deal with various multi-dimensional arrays in python. It is a general purpose array processing package.

Numpy's main purpose is to deal with multidimensional homogeneous array. It has tools ranging from array creation to its handling. It makes it easier to create a n dimensional array just by using np.zeros() or handle its contents using various other methods such as replace, arrange, random, save, load it also helps I array processing using methods like sum, mean, std, max, min, all, etc

Array created with numpy also behave differently then arrays created normally when they are operated upon using operators such as +,-,*,/.

All the above qualities and services offered by numpy array makes it highly suitable for our purpose of handling data. Data manipulation occurring in arrays while performing various operations need to give the desired results while predicting outputs require such high operational capabilities.

2. pandas : it is the most popular python library used for data analysis. It provides highly optimized performance with back-end source code purely written in C or python.

Data in python can be analysed with 2 ways

- Series
- Dataframes

Series is one dimensional array defined in pandas used to store any data type.

Dataframes are two-dimensional data structure used in python to store data consisting of rows and columns.

Pandas dataframe is used extensively in this project to use datasets required for training and testing the algorithms. Dataframes makes it easier to work with attributes and results. Several of

its inbuilt functions such as replace were used in our project for data manipulation and preprocessing.

1. sklearn: Sklearn is an open source python library with implements a huge range of machine-learning, pre-processing, cross-validation and visualization algorithms. It features various simple and efficient tools for data mining and data processing. It features various classification, regression and clustering algorithm such as support vector machine, random forest classifier, decision tree, gaussian naïve-Bayes, KNN to name a few.

Random Forest

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

This algorithm works on 4 basic steps –

1. It chooses random data samples from dataset.
2. It constructs decision trees for every sample dataset chosen.
3. At this step every predicted result will be compiled and voted on.
4. At last most voted prediction will be selected and be presented as result of classification.

Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Conclusion

Random forest is a great algorithm to train early in the model development process, to see how it performs. Its simplicity makes building a “bad” random forest a tough proposition.

The algorithm is also a great choice for anyone who needs to develop a model quickly. On top of that, it provides a pretty good indicator of the importance it assigns to your features.

Random forests are also very hard to beat performance wise. Of course, you can probably always find a model that can perform better, like a neural network for example, but these usually take more time to develop, though they can handle a lot of different feature types, like binary, categorical and numerical.

Overall, random forest is a (mostly) fast, simple and flexible tool, but not without some limitations.

Insurance cost prediction can be made quiet easy and correct using the Random Forest Algorithm and Linear Regression.

References

- www.geeksforgeeks.com
- www.github.com
- https://stackoverflow.com/
- https://scikit-learn.org/stable/