BIG DATA TECHNOLOGY

MINI Project

Medical insurance cost predication

project by:-

PC 25 Katyayini Gupta

PC 26 Swati Singh

PH 16 Aaditya Singh

PD 19 Niyati Dubey

PD 26 Atharva Lakkas

Abstract

This paper represents a machine learning-based health insurance prediction system. Recently, many attempts have been made to solve this problem, as after Covid-19 pandemic, health insurance has become one of the most prominent areas of research. We have used the USA's medical cost personal dataset from kaggle, having 1338 entries. Features in the dataset that are used for the prediction of insurance cost include: Age, Gender, BMI, Smoking Habit, number of children etc. We used linear regression and also determined the relation between price and these features. We trained the system using a 70-30 split and achieved an accuracy of 81.3%

I. INTRODUCTION

we are on a planet full of threats and uncertainty. people, households, companies, properties, and property are exposed to different risk forms. and the risk levels can vary. these dangers contain the risk of death, health, and property loss or assets. life and wellbeing are the greatest parts of people's lives. but, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to reimburse them. insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks[.concerning the value of insurance in the lives of individuals, it becomes important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it. various variables estimates these charges. each factor of these is important. if any factor is omitted when the amounts are computed, the policy changes overall. it therefore critical that these tasks are performed with high accuracy. as human mistakes are could occur, insurers use people with experience in this area. they also use different tools to calculate the insurance premium. ml is beneficial here. ml may generalize the effort or method to formulate the policy.

since there are many independent variables used to calculate the dependent(target) variable. for this study, the dataset for cost of health insurance is used [2]. preprocessing of the dataset done first. then we trained regression models with training data and finally evaluated these models based on testing data. in this paper, we used several models of regression, for example, multiple linear regression, generalized additive model, svm, rf, decision tree (cart), xgboost, k-nearest neighbors, stochastic gradient boosting,

and deep neural network. it is found that the stochastic gradient boosting provides the highest accuracy with an r-squared value of 85. 8295. the key reason for this study is to include a new way of estimating insurance costs.

II. DATASET

To create the claim cost model predictor, we obtained the data set through the Kaggle site (2). The data set includes seven attributes see table 1; the data set is separated into two-part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model. the following table shows the Description of the Dataset. **Table I. Dataset overview**

name	Description
age	Age of the client
BMI	body mass index
The Number of Kids	number of children the client have
gender	Male / Female
smoker	whether a client is a smoker or not
region	where the client lives southwest, southeast, northwest or northeast
Charges(target variable)	Medical Cost the client pay

III. DATA PREPROCESSING

The dataset includes seven variables, as shown in table 1. every one of these attributes has some contribution to estimate the cost of the insurance, which is our dependent variable. In this stage, the data is scrutinized and updated properly to efficiently apply the data to the ML algorithms. First of all, all unknown values are cleaned. The unknown numerical values are replaced with the mean. The target variable (charges) would then be examined.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1122	4740	9382	13270	16640	63770

Table II

IV. METHODOLOGY

1. Data Collection

Importing python libraries for data collection like pandas, numpy, matplotlib, seabon, sklearn etc. numpy libraries provides numpy arrays. Pandas provides data frames so that processing is easy. Matplotlib plots graphs. seabon is data visualization library and sklearn helps in regression. Loading the CSV file into to pandas dataframe here insurance_dataset. There are about 1338 rows and 7 columns and about 1338 entries.

Getting some information about the dataset and checking for missing values. All values are not null so no need for processing the data.

2. DATASET USED

The primary source of data for this project was from Kaggle user Dmarco. The dataset is comprised of 1338 records with 6 attributes. Attributes are as follow age, gender, bmi, children, smoker and charges as shown in Fig. 1. The data was in structured format and was stores in a csv file.

Dataset is not suited for the regression to take place directly. So cleaning of dataset becomes important for using the data under various regression algorithms.

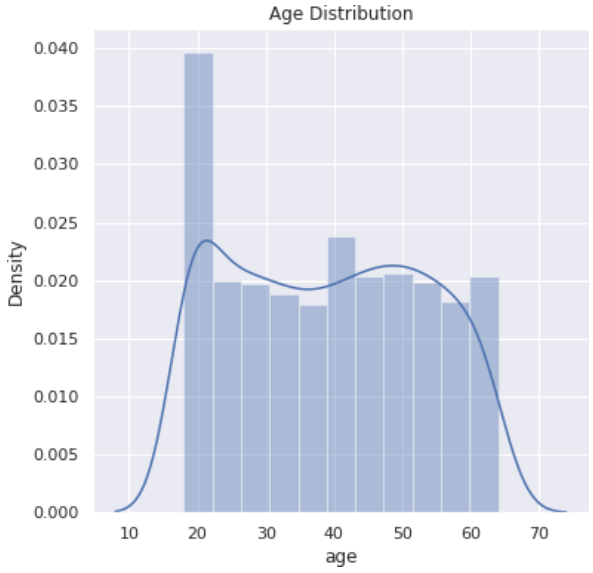
In a dataset not every attribute has an impact on the prediction. Whereas some attributes even decline the accuracy, so it becomes necessary to remove these attributes from the features of the code.

Removing such attributes not only help in improving accuracy but also the overall performance and speed.

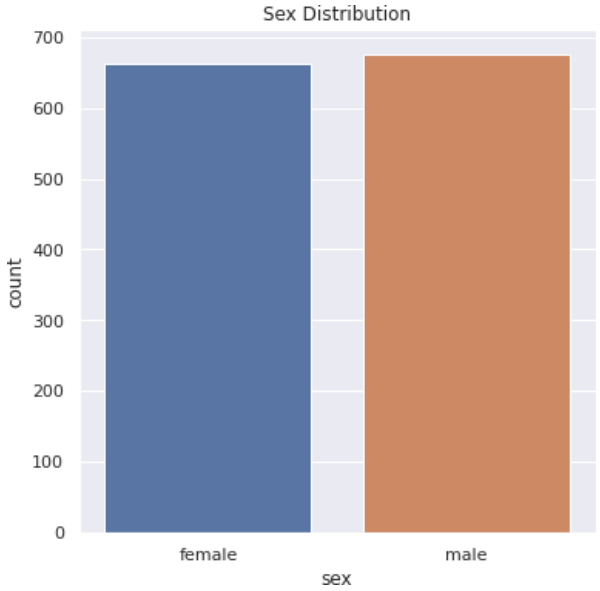
In health insurance many factors such as pre-existing body condition, family medical history, Body Mass Index (BMI), marital status, location, past insurances etc affects the amount. According to our dataset, age and smoking status has the maximum impact on the amount prediction with smoker being the one attribute with maximum effect. Children attribute had almost no effect on the prediction, therefore this attribute was removed from the input to the regression model to support better computation in less time.

3. Data Analysis

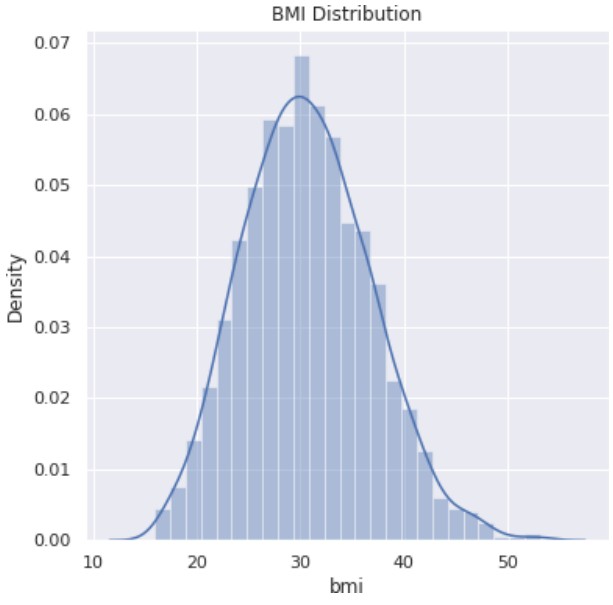
Statistical measures of data set is provided. In which only numerical fields are accepted because other fields are category fields.



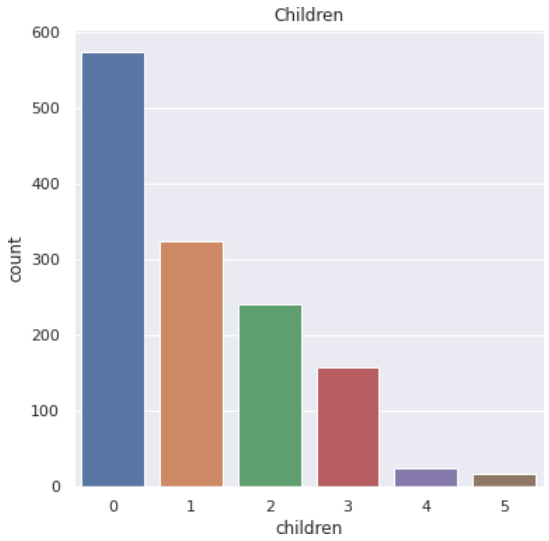
Age Distribution graph shows that the age number of 20 has the highest medical insurance density.



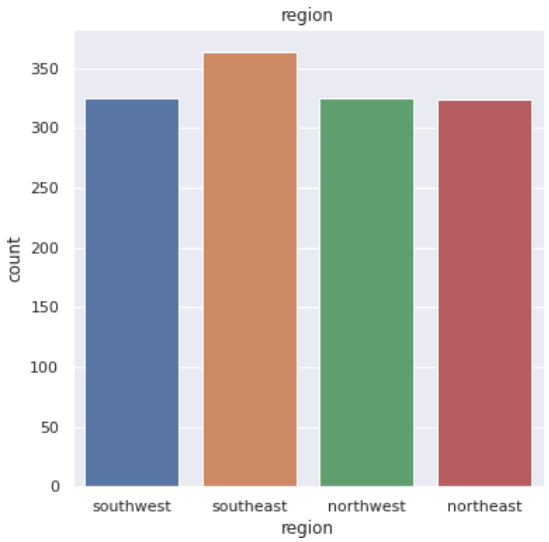
Sex distribution shows that the males have slightly higher medical insurance count than females.



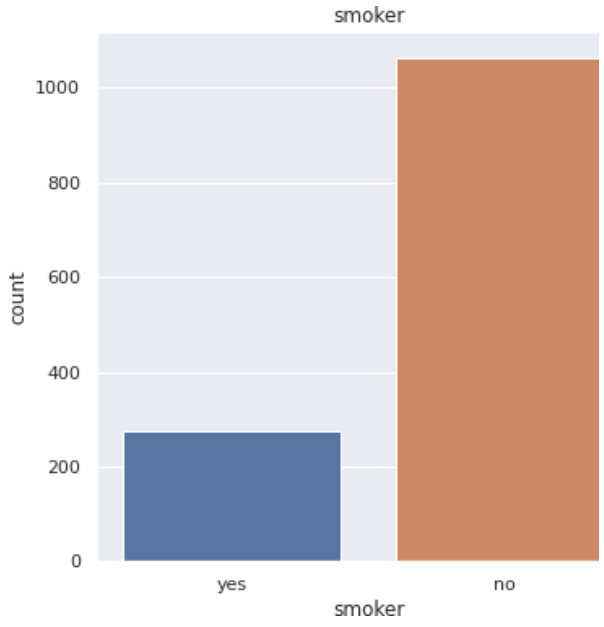
BMI distribution shows that people with bmi of 30 have the highest of the density of the medical insurance count.



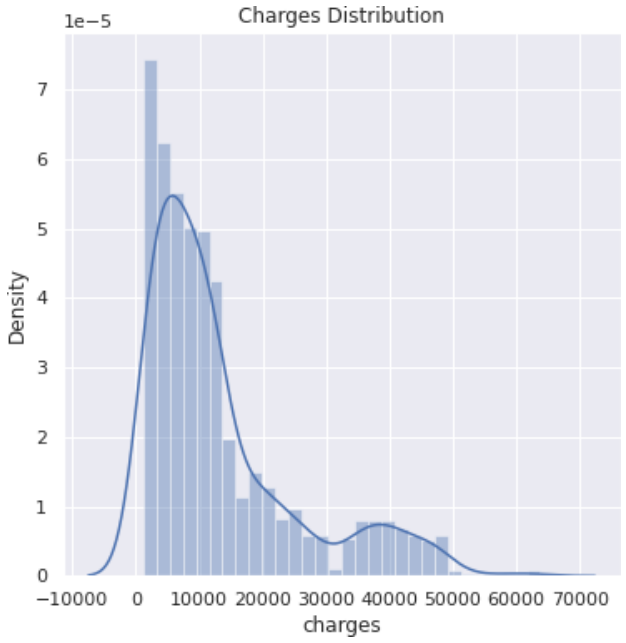
This graph shows that the adults with 0 children have the highest count of medical insurance count in USA.



The southeast region has higher compared to the others followed by southwest northwest and northeast respectively.



This graph shows that the adults who smokes have low rate of medical insurance than the one who does in the USA



The chargers were higher at first then were decreasing

V. RELATED WORKS

A. LITERATURE REVIEW

In this section, research efforts from the exploration of information and machine learning techniques are discussed. Several papers have discussed the issue of claim prediction. Jessica Pesantez-Narvaez suggested, "Predicting motor insurance claims using telematics data" in 2019. This research compared the performance of logistic regression and XGBoost techniques to forecast the presence of accident claims by a small number and results showed that because of its interpretability and strong predictability[3], logistic regression is an effective model than XGBoost., this system takes pictures of the damaged car as inputs and produces relevant details, such as costs of repair to decide on the amount of insurance claim and locations of damage. Thus the predicted car insurance claim was not taken into account in the present analysis but was focussed on calculating repair costs[4]. Oskar Sucki 2019, The purpose of this research is to study the prediction of churn. Random forests were considered to be the best model (74 percent accuracies). In some fields, the dataset had missing values. Following an analysis of the distributions, the decision has been taken to substitute the missing variables with additional attributes suggesting that this data does not exist [5]. This is permitted only if the data is absolutely randomly lost, and so the missing data mechanism by which the appropriate approach to data processing is decided has first to be established[6][7 paper, the exactness of XGBoost is applied to predict statements. Compare the output with the performance of XGBoost, a collection of techniques e.g., AdaBoost, Random Forest, Neural Network. XGBoost offers better Gini structured accuracy. Using publicly accessible Porto Seguro to Kaggle datasets. The dataset includes huge quantities of NaN values but this paper manages missing values by medium and median replacement. However, these simple, unprincipled methods have also proven to be biased[7]. They, therefore, concentrate on exploring the methods ML that are highly appropriate for thproblems of several missing values, such as XGboost, in 2018, Three classifiers have been developed in this study to predict and estimate fraudulent claims and a percentage of premiums for the various customers based upon their personal and financial data. For classification, the algorithms Random Forest, J48, and Naïve Bayes are chosen. The findings show that Random Forest exceeds the remaining techniques depending on the synthetic dataset. This paper therefore does not cover insurance claim forecasts, but rather focuses on false claims [9]. The above previous works did not consider both predicted the cost or claim severity, they only make a classification for the issues of claims (whether or not a claim was filed for that policyholder) in this study we focus on advanced statistical methods and machine learning algorithms and deep neural network for predict the cost of health insurance

B. Regression

The regression analysis is a predictive method that explores the relationship between a dependent (target) and the independent variable(s) (predictor). This technology is used to forecasting, estimate model time series, and find the causal effect relationship among the variables. In this analysis, for example, I want to analyze the relationship between insurance cost (target variable) and six independent variables based on (age, BMI, child number, individual living area, or sex and

whether the customer is a smoking person).on the basis of a regression. The regression analysis estimates the relationship between two or more variables, as stated previously. I used different regression models to estimate health insurance costs on the basis of six independent variables, and by using this regression, we can forecast future health insurance fees based on current and past data. There are several advantages of using regression analysis as follows: -It demonstrates the essential relationships between the dependent and independent variables. -It shows the effect intensity on the dependent variable of several independent variables. Analysis of regression also helps one to compare the results of measured variables at various scales, such as independent variable and dependent variable effects. These advantages allow market researchers, data analysts, and data scientists to remove and determine the best range of variables for predictive model

I. REGRESSION MODELS

1) Multiple Linear Regression. In practice, we often have more than one predictor. For example, with the data set used in this study, we may wish to understand if independent variables (6 independent variables), (linearly) related to the dependent variable (charges). this is referred to as the multiple linear regression (MLR) model [10]. An MLR model with t\ independent features , , ..., , and Y results can be calculated as in the following equation In the above equation, u is the residual regression while a is the weight of each independent variable or parameter assigned.

2) Generalized Additive Model (GAM) Generalized additive models are incorporated into the actuary toolkit to deal flexibly with continuous functionality. The continuous features in this setting insert the model into a semi-parametric additive predictor. The impact of the policyholder's age, vehicle power or amount insured may be modeled by GAMs in property and casualty insurance. GAMs also allow actuaries to evaluate geographical risk variances, taking into account the potential interaction of continuous characteristics. Other experiences in the data usually include age, power and gender, and age in engine insurance. You can also be caught by GAMs.[11]

3) Random Forest Random forests reflect a shift to the bagged decision trees that create a broad number of decorrelated trees so that predictive efficiency can be improved further. They are a very popular 'off-the-box' or off-the-shelf' learning algorithm, with good predictive performance and relatively few hyperparameters. There are several implementations of random forests that exist, but the Leo Breiman algorithm (Breiman 2001)[12] is now largely authoritative. .Random forests create an average predictive value as a result throughout the regression of individual trees. Random forests resolve to overfit [10]. As in the following equation , a random model for forest regressors can be expressed. where g is the final model that is the sum of all models. Each model f(x) is a decision tree.

4) XGBoost. Recently a new ensemble learning software named XGBoost has been proposed[13]. Which is a new tree-enhancing model that provides effective out-of-core learning and sparse memory. XGBoost is therefore a supervised learning algorithm, which would be extremely useful for argument prediction issues with broad training data and missing values. Missing values can still not be managed by the most popular approaches, such as

random forests and neural networks. Methods require additional frameworks to manage the missing values. The power of XGBoost improves the use of the tool in many other applications. For example, in direct-diffuse solar separation, Aler et al. [14] Developed two versions of XGBoost. The first one is an indirect model, which uses XGBoost to learn solar radiation separation models from various literature sources in a data set from traditional level 1 instruction models. Another model is a direct model that straightforwardly suits XGboost in a dataset. An Additional case is [15], which uses XGBoost to recommend things to a user using functions derived from the pair of users using complicated feature engineering in the recommendation framework. In this study, we analyze XGBoost as a predictor model for the medical insurance cost

5) Support Vector Machine SVMs can be generalized to problems with regression (i.e., when the outcome is continuous such as our target variable in our study). Essentially, SVMs are seeking a hyperplane in an extended function space that usually results in a nonlinear decision limit with strong generalization efficiency in the original feature space. Specific functions called kernel functions are used to build this extended, separated functionality.

6) K-Nearest Neighbors K-NN is a very simple predictive model that predicts values on the basis of their "likelihoods" from other values. Contrary to most other machine learning approaches, KNN is dependent on memory and cannot be summed up as a closed algorithm. This implies that the training data are required during operations and forecasts are produced immediately from the training data relations. KNNs are additionally identified as lazy learning[16] and can also be inefficient computationally. Nevertheless, KNNs have succeeded in several market problems [17][18].

7) Stochastic GBMs Gradient boosting machines (GBMs) are an extremely common ML regression model. Whereas random forests construct a group of independent deep trees, GBMs construct a set of shallow trees Each tree learns and develops compared to the prior one. International Journal of Innovative

While shallow trees are feeble forecasting predictive models, they can be "boosted" to create a strong committee, which often is difficult to tackle with other algorithms if properly tuned. A significant observation from Breiman [19][20] was that the training of algorithms on a random training subsample provided more reductions in the tree correlation and thus enhanced predictive accuracy. The same logic was used by Friedman (2002)[21] and the boosting algorithm upgraded accordingly. This procedure is called a stochastic gradient boosting.

VI. IMPLEMENTATION

The objective of the study is to predictive the insurance cost based on age, BMI, child number, the region of the person living, sex, and whether a client is smoking or not. These features contribute to our target variable prediction of insurance costs. For the measurement of the cost of insurance, several regression models are applied in this study. The dataset is split into two sections. One part for model training and the other part for model evaluation or testing. In this study, the data set is separated into two-part the first part is called training data and the second called test data, training data makes up about 80 percent of the total data used, and the rest for test data. Every

one of these models is trained with the training data part and then evaluated with the test data[26–28]. For this study, R x64 4.0.2 is used for applying these models. We used two main libraries are CART and Keras for ML and deep learning models. And we used Mean absolute error (MAE), root mean squared error (RMSE) and R-squared As a standard for evaluating these models

The Mean Absolute Error (MAE) is the difference between the original and forecast values obtained by averaging the absolute difference over the data set.

$$MAE=\frac{1}{N}\sum_{n=1}^{N}|\check{Y}-Y|$$

The RMSE of the disparity between the expected values and the real values is determined as the square root. For an accurate forecast, the RMSE must be low so there would be less variance among the expected values and the real values.

$$RMSE=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\check{Y}-Y)^2}$$

Where N = Number of overall observations, $\check{Y}$ = expected insurance fee values, y = real insurance fee values.

The R-squared is often called the coefficient of decision. The proportion of variance is estimated from the independent variables in the dependent variable.

$$R\text{-squared}=\frac{\text{Explained variance}}{\text{Total variance}}$$

The more R-squared , the better the model output. , and indicates that the model deviates less from real values. A R-squared score of 1 indicates that it suits perfectly.

To evaluate the performance of various machine learning algorithms (Multiple Linear Regression, Generalized Additive Model, SVM, CART, RF, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network All of these models are trained on the basis of training data and tested on test data. Mean absolute error, RMSE, and R-squared for each of these models are measured.

And Table IV displays the results.

Algorithm used	MAE	RMSE	R-squared
Stochastic Gradient Boosting	0.17448	0.380189	0.858295
XGBoost	0.213859	0.382509	0.853653
Random Forest Regressor	0.215625	0.388319	0.849299
Support Vector Machine	0.234765	0.394699	0.842307
Decision tree(CART)	0.240118	0.403336	0.833493
DNN	0.254768	0.421432	0.809799
Generalized Additive Model	0.289473	0.445469	0.757636
Multiple Linear Regression	0.28636	0.449725	0.755813
k-Nearest Neighbors	0.574117	0.766835	0.318513

VII. CONCLUSION

The research uses various machine learning regression models and deep neural networks to forecast charges of

health insurance based on specific attributes, on medical cost personal data set from Kaggle.com. The findings are summarized in Table IV. shows that Stochastic Gradient Boosting offers the best efficiency, with an RMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.82. Stochastic gradient boosting can therefore be used in the estimation of insurance costs with better performance than other regression models. Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. Machine learning can significantly minimize these individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

CONCLUSION & FUTURE SCOPE

TO find the insinsurance cost premium various factors were used and their effect on predicted amount was examined. It was observed that a person's age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features.

The effect of various independent variables on the premium amount was also checked. The attributes also in combination were checked for better accuracy results.

Premium amount prediction focuses on persons own health rather than other company's insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount.

Refrences

1. Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big

data analytics with health insurance in India. In 2016 International

Conference on Innovation and Challenges in Cyber Security

(ICICCS-INBUSH) (pp. 64-69). IEEE.

2. Kaggle Medical Cost Personal Datasets. Kaggle Inc.

https://www.kaggle.com/mirichoi0218/insurance.

3. Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting

motor insurance claims using telematics data—XGBoost versus

logistic regression. Risks, 7(2), 70

4. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019,

September). Automating Car Insurance Claims Using Deep Learning

Techniques. In 2019 IEEE Fifth International Conference on

Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

5. Stucki, O. (2019). Predicting the customer churn with machine

learning methods: case: private insurance customer data.

6. Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P.,

Kenward, M. G., ... & Carpenter, J. R.

(2009). Multiple imputation for

missing data in epidemiological and clinical research: potential and

pitfalls. Bmj, 338.

7. Van Buuren, S. (2018). Flexible imputation of missing data. CRC

press.

8. Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for

insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2).

9. Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent

claims in automobile insurance. In 2018 Second International

Conference on Inventive Communication and Computational

Technologies (ICICCT) (pp. 1338-1343). IEEE.

10. Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance

comparison of multiple linear regression, random forest and artificial

neural network by using photovoltaic and atmospheric data. In 2017

14th International Conference on Engineering of Modern Electric

Systems (EMES) (pp. 1-4). IEEE.

11. Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019).

Effective Statistical Learning Methods for Actuaries I: GLMs and

Extensions. 10.1007/978-3-030-25820-7.

12. Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1).

Springer: 5–32.

13. Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting

system 22nd ACM SIGKDD Int. In Conf. on Knowledge Discovery

and Data Mining.

14. Aler, R., Galván, I.M., Ruiz-Arias, J.A., Gueymard, C.A. (2017).

Improving the separation of direct and diffuse solar radiation

components using machine learning by gradient boosting. In Solar

Energy vol. 150, pp. 558-569.

15. Volkovs, M., Yu, G. W., & Poutanen, T. (2017). Content-based

neighbor models for cold start in recommender systems.

In Proceedings of the Recommender Systems Challenge 2017 (pp.

1-6).

16. Cunningham, Padraig, and Sarah Jane Delany. 2007. ―K-Nearest

Neighbour Classifiers.‖ Multiple Classifier Systems 34 (8). Springer

New York, NY, USA: 1–17

17. Jiang, Shengyi, Guansong Pang, Meiling Wu, and Limin Kuang. 2012.

―An Improved K-Nearest-Neighbor Algorithm for Text

Categorization.‖ Expert Systems with Applications 39 (1). Elsevier:

1503 9.

18. Mccord, Michael, and M Chuah. 2011. ―Spam Detection on Twitter

Using Traditional Classifiers.‖ In International Conference on

Autonomic and Trusted Computing, 175–

86. Springer.

19. Breiman, L. (1996). Bagging predictors. Machine learning, 24(2),

123-140

20. Breiman, Leo, and others. 2001. ―Statistical Modeling: The Two

Cultures (with Comments and a Rejoinder by the Author).‖ Statistical

Science 16 (3). Institute of Mathematical Statistics: 199–231.

21. Friedman. 2002. ―Stochastic Gradient Boosting.‖ Computational

Statistics & Data Analysis 38 (4). Elsevier: 367–78.

22. Sabbeh, S. F. (2018). Machine-learning techniques for customer

retention: A comparative study. International Journal of Advanced

Computer Science and Applications, 9(2).

23. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of

machine learning. MIT press.

24. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications

for classification and prediction. Shanghai archives of

psychiatry, 27(2), 130.

25. Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep

learning (Vol. 1, No. 2). Cambridge: MIT press.

26. Kansara, Dhvani & Singh, Rashika & Sanghvi, Deep & Kanani, Pratik.

(2018). Improving Accuracy of Real Estate Valuation Using Stacked

Regression. Int. J. Eng. Dev. Res. (IJEDR)

6(3), 571–577 (2018)