EAA e-Conference on
Data Science & Data Ethics

14 May 2024

Prediction of Health Care Insurance Expenses using Machine Learning and Artificial Neural Network

Liana Barsoumian / Dr Re-Mi Hage

Notre Dame University - Louaize

AGENDA

Introduction

Literature Review

Methodology

Modeling Results

Conclusion

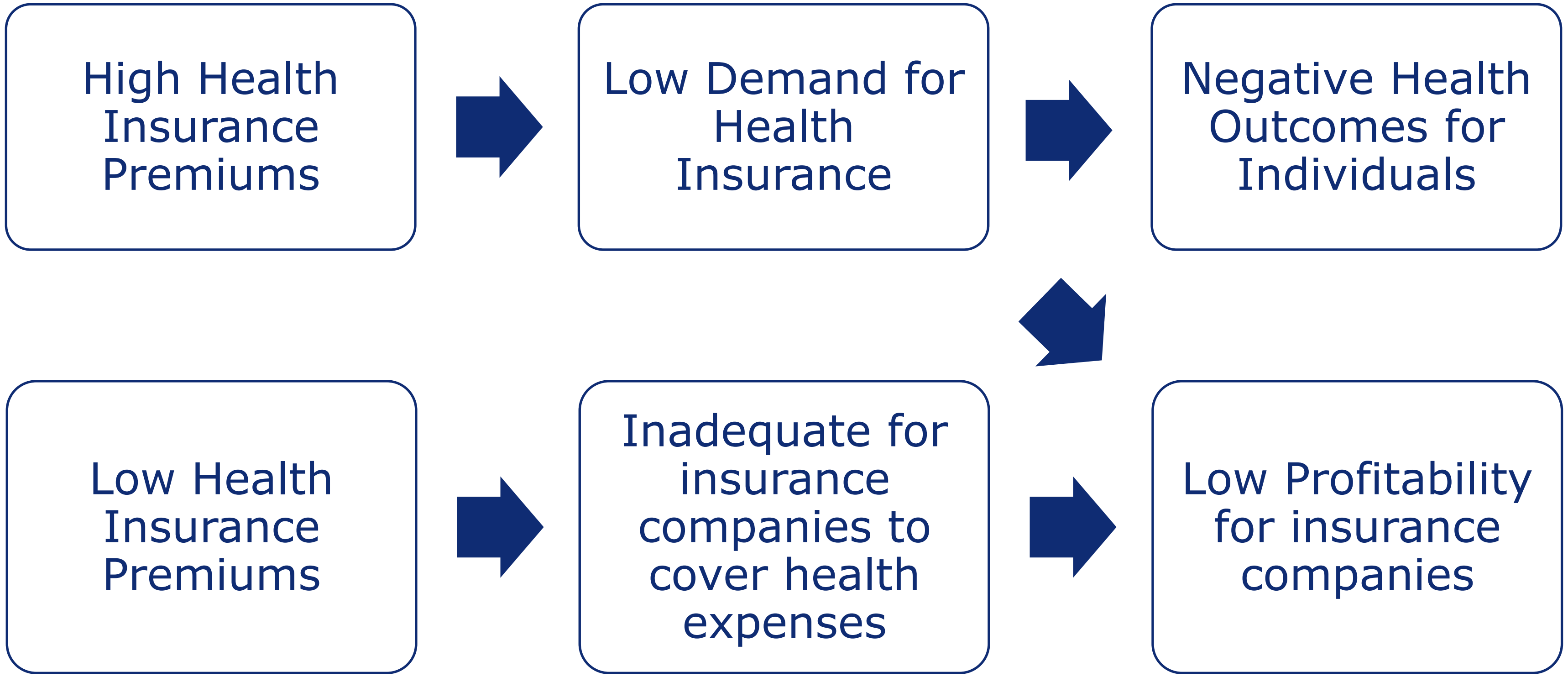
Future Work

Questions

Health Insurance Premiums:

- ❑ Provide protection against high-cost medical treatments.
- ❑ Vary based on several medical and demographic information of the individual.
- ❑ Are continuously on the rise, resulting in a demand for more affordable options.

IMPORTANCE OF AFFORDABLE AND ACCURATE HEALTHCARE



ROLE OF AI AND MACHINE LEARNING IN HEALTHCARE INSURANCE

- Analyze large amounts of data at high computational speeds.
- Assist in predicting healthcare expenses to modify premiums accordingly.
- Detect individuals with a high risk of developing chronic illnesses.
- Detect and prevent fraud.

AGENDA

- Introduction
- **Literature Review**
- Methodology
- Modeling Results
- Conclusion
- Future Work
- Questions

THE MOST IMPORTANT MACHINE LEARNING MODELS STUDIED IN RECENT LITERATURE PAPERS (SINCE 2018)

Machine Learning Models	Accuracy
Multiple Linear Regression	75-76%
Generalized Additive Model	75%
Support Vector Machine	83-84%
Random Forest	**84-85%**
CART	82-83%
XGBoost	**84-85%**
K-NN	31-32%
Gradient Boosting	**85-86%**
Deep Neural Network	80%
Artificial Neural Network	**92.70%**

european actuarial academy

Data - Ethics - Actuary

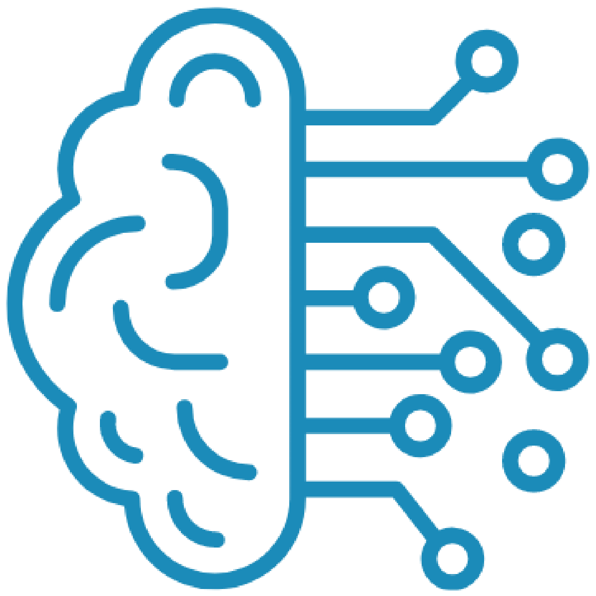
AGENDA

- Introduction
- Literature Review
- **Methodology**
- Modeling Results
- Conclusion
- Future Work
- Questions

METHODOLOGY

The models implemented in this study are:

- ❑ Regression Decision Tree
- ❑ Gradient Boosting Machine
- ❑ XGBoost
- ❑ Multiple Linear Regression
- ❑ Feed-Forward Artificial Neural Network

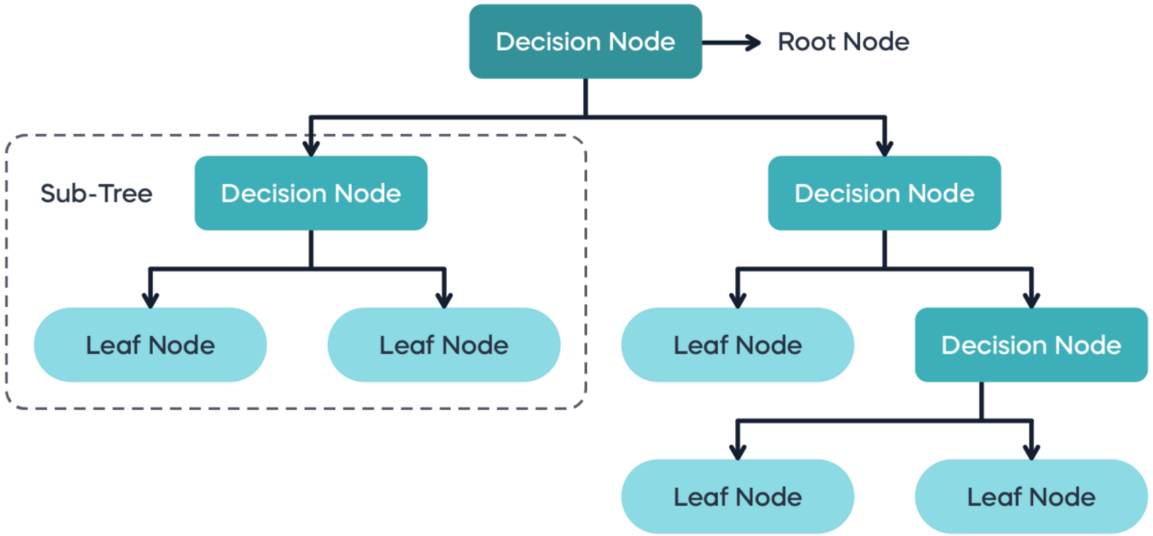


REGRESSION DECISION TREE

- It is supervised learning method that can be used for solving both classification and regression problems.
- It requires the “Recursive Binary Splitting” approach.

$$R_1(j,s) = \{X|X_j < s\} \; and \; R_2(j,s) = \{X|X_j \geq s\}$$

- It minimizes the Sum of Squared Errors (SSE) at each stage of the tree.
- It might sometimes overfit the dataset and perform poorly on new, unseen data.
- It can be pruned to improve the accuracy of the tree.



REGRESSION DECISION TREE

Pruning a Decision Tree

Pre-Pruning

Sets a limit to the:

- Depth of the decision tree
- Number of samples to split a node

Post-Pruning

- Builds the full decision tree
- Removes nodes that do not significantly improve accuracy
- $R_\alpha(T) = R(T) + \alpha|T|$, where α is the cost-complexity parameter

GRADIENT BOOSTING MACHINE

- An effective machine learning method for both classification and regression issues.
- Can handle both numerical and categorical data.
- Combines various decision trees (weak learners) to produce a stronger model.
- Builds up in a sequential manner, where each learner corrects the errors of the preceding one.
- Predicts values of the form: $\hat{y} = F(x)$ by minimizing the Mean Squared Error (MSE).
- Uses a gradient descent optimization algorithm to determine the weights of the weak learners.

GRADIENT BOOSTING MACHINE

Number of Trees

Optimal number of trees to reduce overfitting.

Depth of Trees

To control the complexity of the tree.

Learning Rate

Smaller values reduce overfitting but increase computational time.

Subsampling

When fewer than 100% of the training observations are used, stochastic gradient descent is applied.

XGBOOST

- A powerful and well-known machine learning technique in the gradient boosting family used for classification, regression, and ranking
- Operates in a manner similar to that of the Gradient Boosting Machine.
- Has a higher computational speed.
- Includes a regularization term to control the model’s complexity and avoid overfitting or underfitting the dataset.

MULTIPLE LINEAR REGRESSION

- One of the most simple and common algorithms used in statistics and machine learning.
- Refers to the relationship between a dependent variable Y and multiple independent variables X

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_k X_{ik} + \varepsilon_i$$

- Applies the method of “Least Squares” to minimize the Sum of Squared Errors (SSE).
- The coefficients are unbiased, consistent, sufficient, and have minimum variance.

MULTIPLE LINEAR REGRESSION

Assessing the Goodness of Fit of the Model

R-Squared

Quantifies the proportion of the total variance in the dependent variable that is explained by the independent variables.

Adjusted R-Squared

Same as R-squared but considers the number of independent variables and the sample size.

F-Test

Assesses the overall statistical significance of the Regression model.

T-Test

Assesses the significance of individual independent variables.

BP (Breusch Pagan) Test

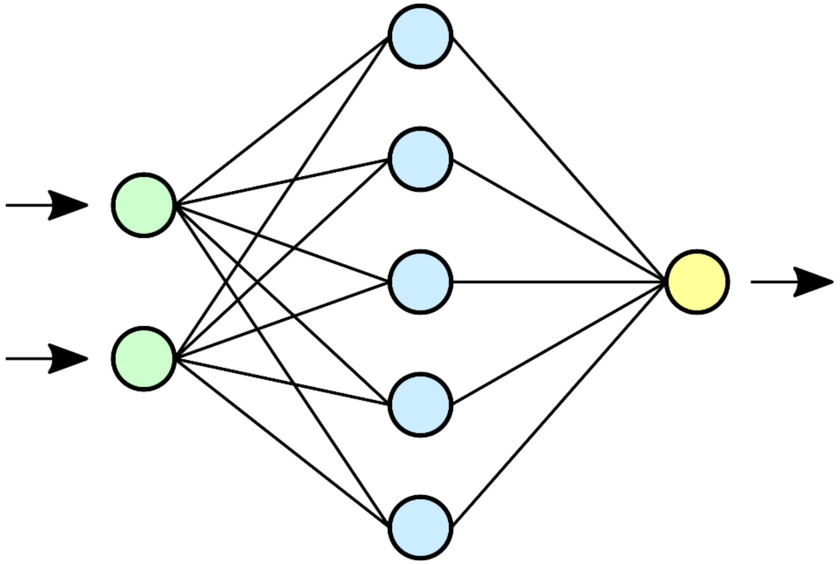
Used to detect whether heteroscedasticity is present (assumption of constant variance).

FEEDFORWARD ARTIFICIAL NEURAL NETWORK

- A collection of algorithms that aims to identify underlying relationships in a group of data using a method that imitates how the human brain functions.
- Consists of millions of artificial neurons.
- Requires large amounts of data to learn new things.
- Is designed to learn from previous outputs and predict future outcomes.
- Possesses high computational skills and speeds.
- Has become increasingly popular in a variety of fields.

FEEDFORWARD ARTIFICIAL NEURAL NETWORK

- It consists of an input layer, one or several hidden layers, and an output layer.
- The input nodes are connected with an activation function to transform them into outputs.
- Each node multiplies the input signal with a weight w_{ij}, characteristic of the connection between nodes i and j of layers to relate the weighted input.



Input is fed to the first layer to be processed → Hidden layer performs numerical calculations → Output layer displays the acquired results

PERFORMANCE METRICS

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

Mean Absolute Percentage Error (MAPE)

AGENDA

- Introduction
- Literature Review
- Methodology
- **Modeling Results**
- Conclusion
- Future Work
- Questions

DATA ANALYSIS

Data Preparation – Excerpt of the dataset

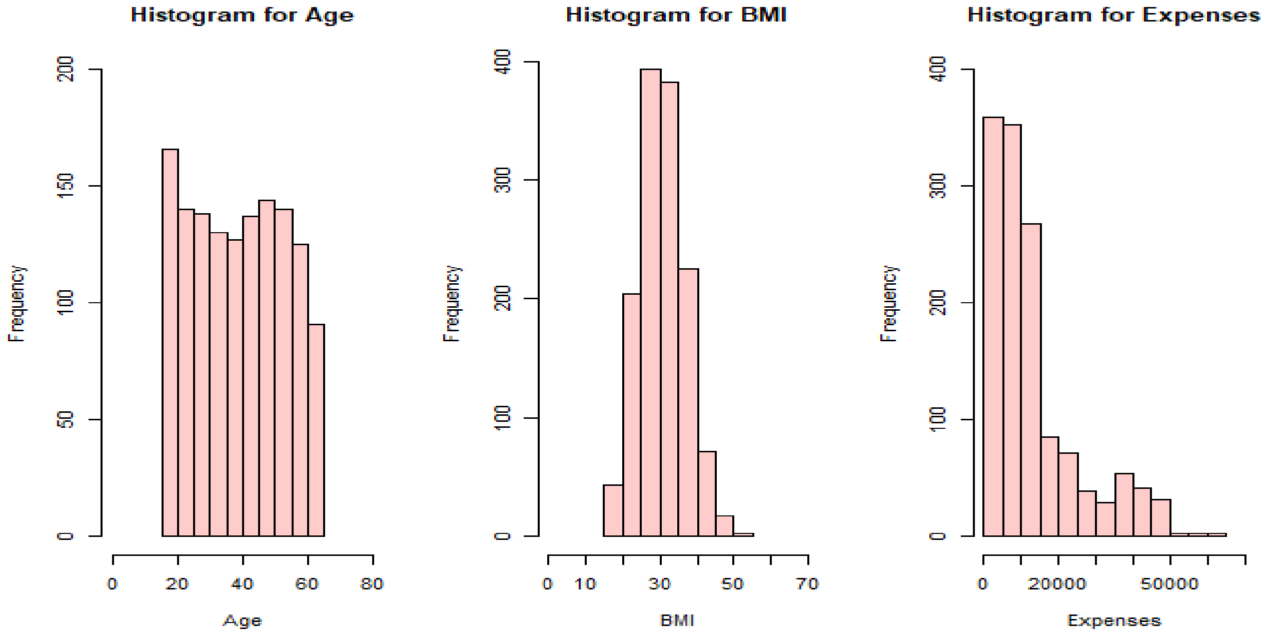
- ❑ Consists of 7 variables: Age, Sex, BMI, Children, Smoker, Region and Expenses.
- ❑ Includes 1,338 observations.
- ❑ Age, BMI and Expenses are considered numerical, whereas Sex, Children, Smoker and Region are considered categorical variables.
- ❑ 70% of the dataset is considered to be training dataset, and the remaining 30% is used as the testing dataset.

Age	Sex	BMI	Children	Smoker	Region	Expenses
19	female	27.9	0	yes	southwest	16,884.92
18	male	33.8	1	no	southeast	1,725.55
28	male	33	3	no	southeast	4,449.46
33	male	22.7	0	no	northwest	21,984.47
32	male	28.9	0	no	northwest	3,866.86
31	female	25.7	0	no	southeast	3,756.62
46	female	33.4	1	no	southeast	8,240.59
37	female	27.7	3	no	northwest	7,281.51
37	male	29.8	2	no	northeast	6,406.41
60	female	25.8	0	no	northwest	28,923.14

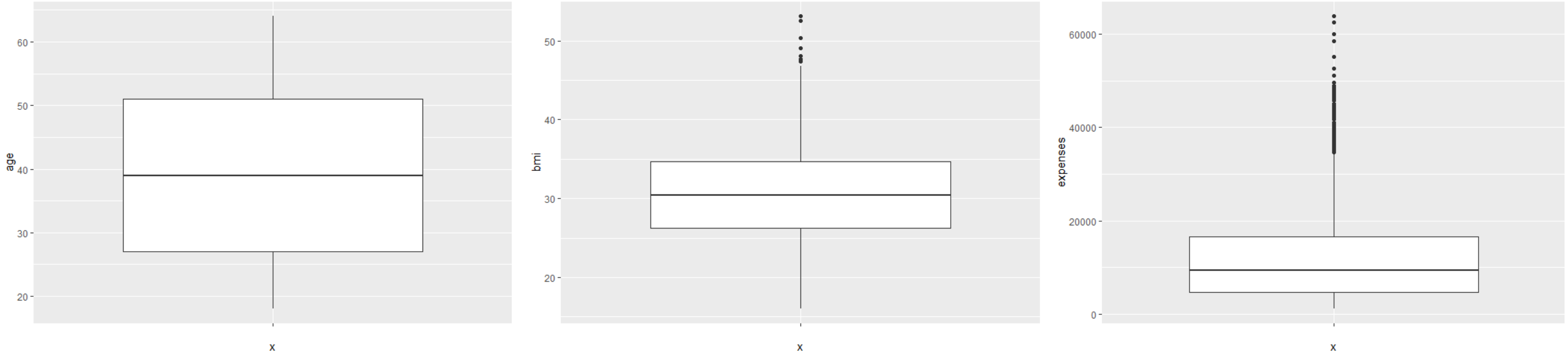
DATA ANALYSIS – UNIVARIATE ANALYSIS

Variables	Min.	Q1	Median	Mean	Q3	Max.	Standard Deviation	Coefficient of Variation
Age	18	27	39	39.21	51	64	14.05	0.36
BMI	16	26.30	30.40	30.67	34.70	53.10	6.10	0.20
Expenses	1,122	4,740	9,382	13,270	16,640	63,770	12,110.01	0.91

Descriptive statistics of the numerical variables of the dataset and their histograms



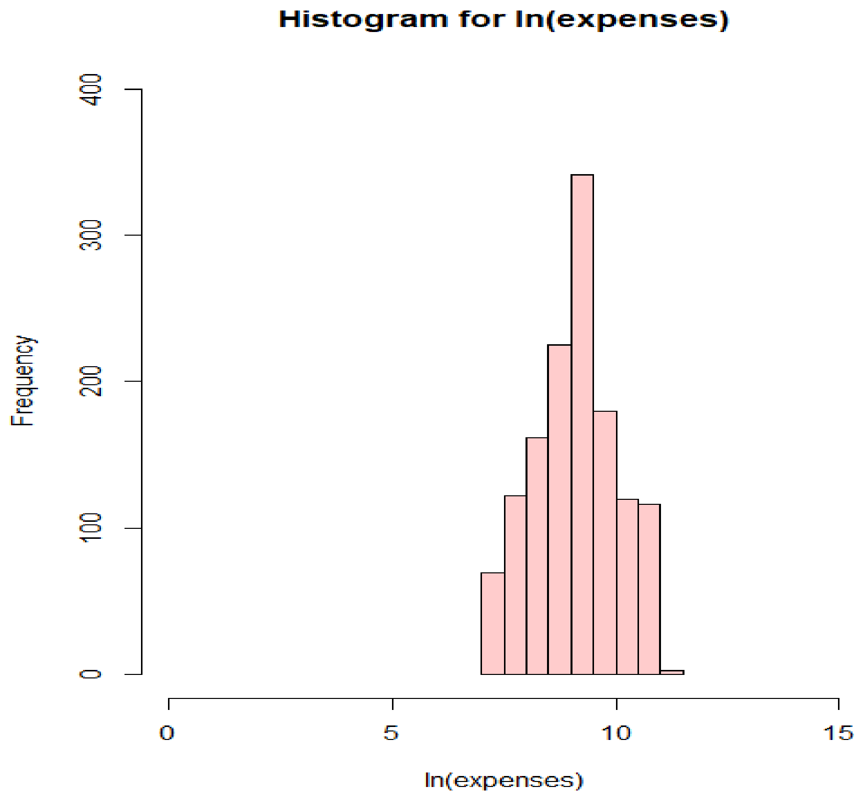
DATA ANALYSIS – UNIVARIATE ANALYSIS



Boxplots of the numerical variables of the dataset

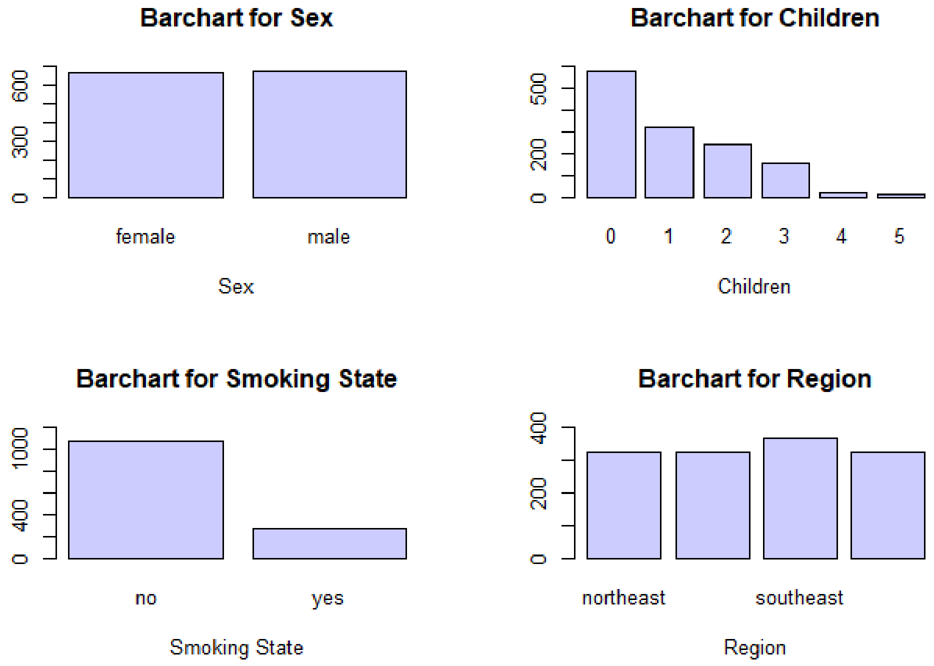
DATA ANALYSIS – UNIVARIATE ANALYSIS

- ❑ Data for Expenses is highly skewed.
- ❑ The high skewness does not meet the assumptions of regression-based models.
- ❑ A logarithmic transformation is applied on Expenses to approximately normalize its distribution.



DATA ANALYSIS – UNIVARIATE ANALYSIS

Variables		Count	Proportion
Sex	Female	662	0.49
	Male	676	0.51
Smoker	Yes	274	0.20
	No	1,064	0.80
Region	Northeast	324	0.242
	Northwest	325	0.243
	Southeast	364	0.272
	Southwest	325	0.243
Children	0	574	0.43
	1	324	0.24
	2	240	0.18
	3	157	0.12
	4	25	0.02
	5	18	0.01



Descriptive statistics of the qualitative variables of the dataset and their bar charts

DATA ANALYSIS – BIVARIATE ANALYSIS

Expenses vs. Numerical Independent Variables

- The Pearson correlation coefficient of Expenses and the numerical variables Age and BMI is calculated.

Variables	Pearson Correlation Coefficient (r)	Covariance
Age and Expenses	0.30	50,874.80
BMI and Expenses	0.19	14,665.15

- A very weak relationship is observed between BMI and Expenses.
- A weak relationship is observed between Age and Expenses.

DATA ANALYSIS – BIVARIATE ANALYSIS

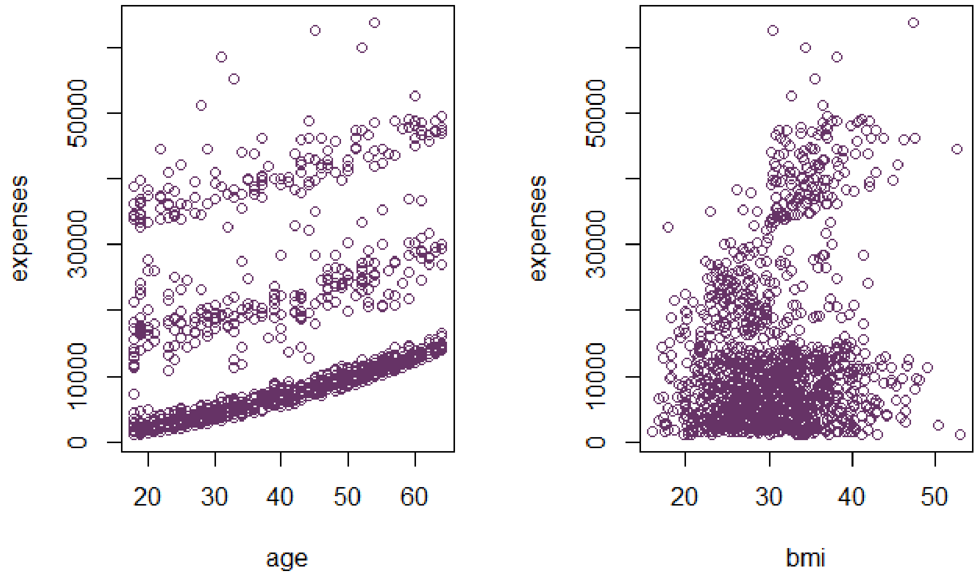
- The Pearson Correlation test can also be used to establish the relationship between two variables.
- *Null Hypothesis*: No relationship exists between the two variables.
- A p-value of 0.05 or below is considered to be statistically significant.

Variables	p-value	95% CI
Age and Expenses	< 2.2e-16	[0.25, 0.35]
BMI and Expenses	2.302e-13	[0.15,0.25]

DATA ANALYSIS – BIVARIATE ANALYSIS

- As Age increases, Expenses also increase.
- High Expenses are associated with high BMIs as well.

Scatterplots of Expenses with respect to Age and BMI



DATA ANALYSIS – BIVARIATE ANALYSIS

Expenses vs. Categorical Independent Variables

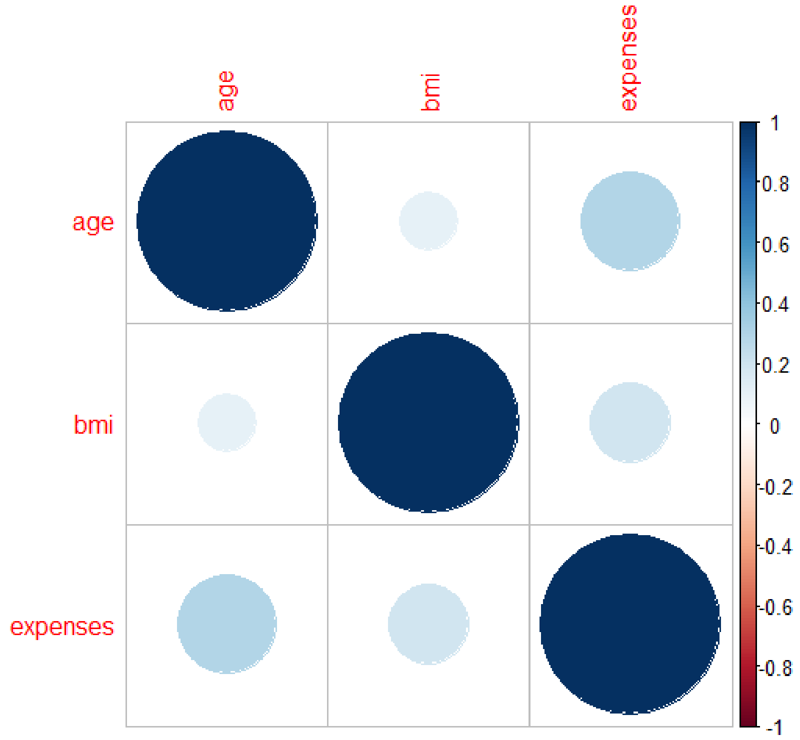
- ANOVA Test to determine if the average of expenses significantly differs between groups of data.
- A p-value of 0.05 or below is considered to be statistically significant.
- All the categorical variables have a significant effect on expenses.

Variables	F-Value	p-value
Sex and Expenses	4.40	0.04
Children and Expenses	6.21	0.01
Smoker and Expenses	2178	<2e-16
Region and Expenses	2.97	0.03

DATA ANALYSIS - MULTIVARIATE ANALYSIS

Numerical Variables

- Dark blue and red shades represent the highest positive and highest negative correlations, respectively.
- The correlation between each pair of numerical variables is shaded light blue, indicating weak positive correlation.



DATA ANALYSIS – MULTIVARIATE ANALYSIS

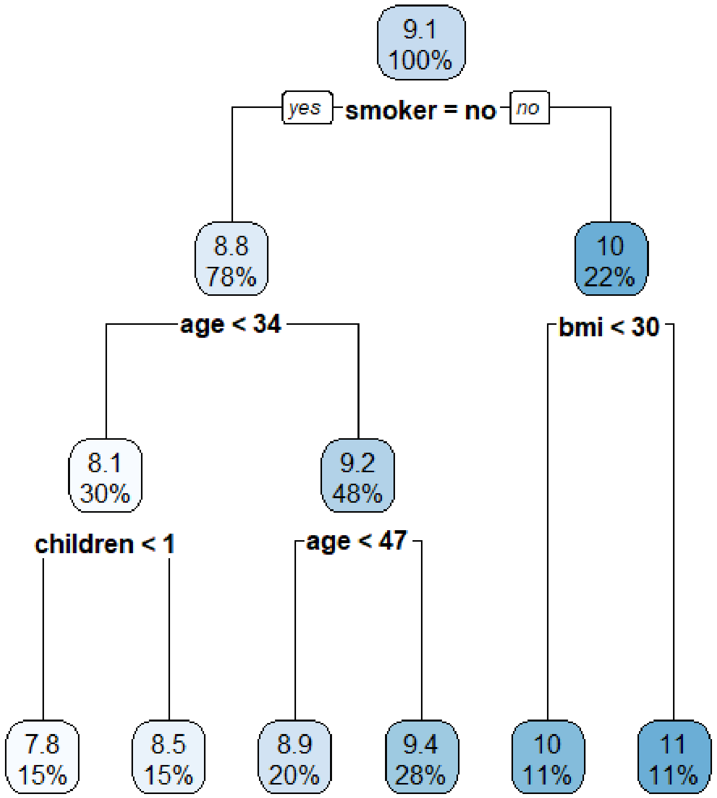
Categorical Variables

- Chi-square test to determine if the two categorical variables are independent.
- A p-value of 0.05 or below shows that the variables are dependent.
- Sex and Smoker are dependent.
- The remaining pairs are independent.

Qualitative Variable Pairs	Chi-square Value	p-value
Sex and Region	0.44	0.93
Sex and Children	0.74	0.98
Sex and Smoker	7.39	0.01
Children and Region	13.77	0.54
Smoker and Region	7.34	0.06
Smoker and Children	6.89	0.23

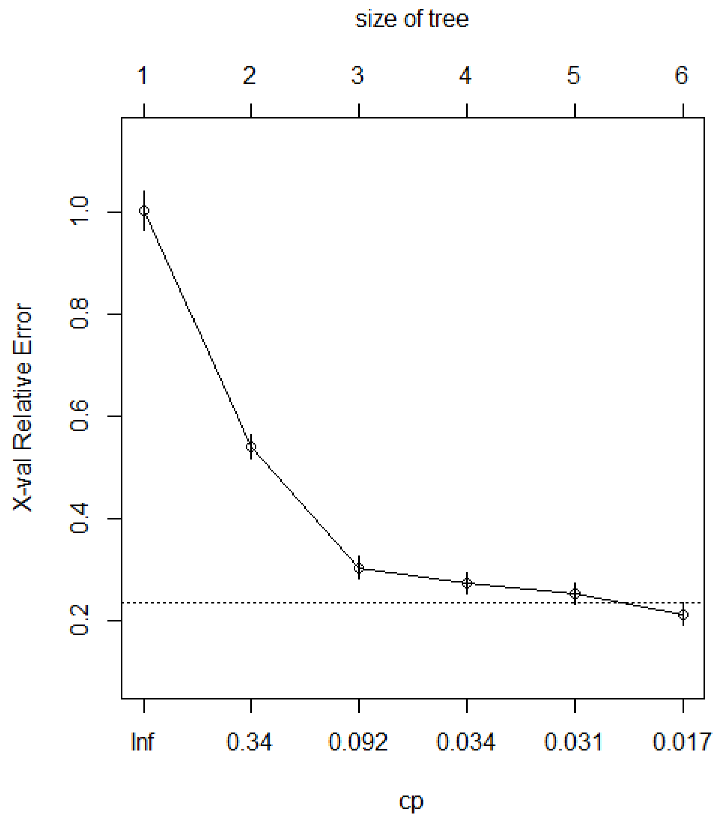
REGRESSION DECISION TREE

- ❑ The first regression decision tree is built using all the independent variables to predict ln(expenses).
- ❑ It displays the percentage of training data that is assigned to each node, and the average amount of ln(expenses) for that branch.
- ❑ The decision tree is partitioning on 4 variables only: Smoker, Age, BMI and Children.



REGRESSION DECISION TREE

- ❑ The algorithm automatically prunes the tree by applying a range of cost-complexity values through a 10-fold cross validation.
- ❑ The cross-validation error is diminishing after 6 trees.
- ❑ The resulting cross-validation error is approximately 0.214.



REGRESSION DECISION TREE

- It is also common to tune the min-split and max-depth of the regression decision tree.
- Testing out different combinations manually can be tedious and time consuming.
- A hyperparameter grid can automatically search over a variety of tuned models.
- The min-split values will range from 5 to 20 and the max-depth values will range from 2 to 7.

Model	Min-Split	Max-Depth	Cost-Complexity Value	Cross-Validation Error
1	8	4	0.01	0.20344
2	17	4	0.01	0.20351
3	9	4	0.01	0.20371
4	9	3	0.01	0.20417
5	13	7	0.01	0.20421

- The cross-validation error is slightly improved from 0.214 to approximately 0.203.

REGRESSION DECISION TREE

- The optimal model is applied to predict on the testing dataset.

RMSE	MAE	MAPE
0.44	0.29	3.22%

- The recorded accuracy is 96.68%.

REGRESSION DECISION TREE

- ❑ Single tree models have a high variance.
- ❑ There are alternate approaches that use the variability of single trees to greatly improve their performance, such as Bootstrap Aggregation (Bagging).
- ❑ Bagging is the process of combining and averaging across numerous models, which lowers variability and overfitting.

Step 1

- Bootstrap samples will be created from the training dataset.

Step 2

- For each sample, an unpruned tree will be trained.

Step 3

- The predictions from each tree will be averaged to obtain an overall average predicted value.

REGRESSION DECISION TREE

- The bagged tree is applied to predict on the testing dataset.

Model	RMSE	MAE	MAPE
Un-Bagged Tree	0.44	0.29	3.22%
Bagged Tree	0.41	0.24	2.70%

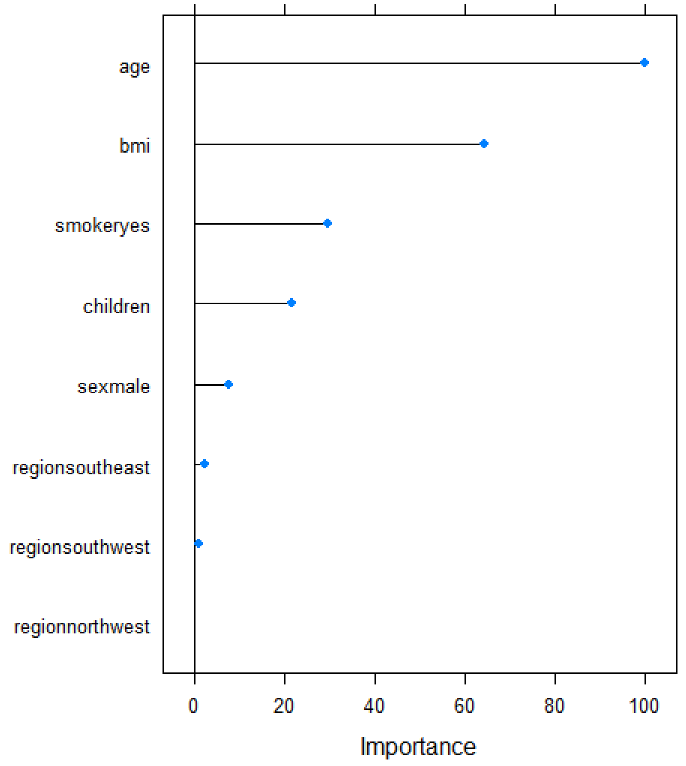
- The RMSE and MAE have decreased compared to the un-bagged tree.
- The recorded accuracy is 97.30%
- There is an improvement in accuracy by 0.62% only.

REGRESSION DECISION TREE

❑ Variable importance can be assessed through bagged decision trees.

❑ The predictors with the greatest average impact on SSE at each split are regarded as the most important.

❑ Age, BMI, Smoker and Children are the most important variables. Sex has some importance, while Region has negligible importance in predicting ln(expenses).

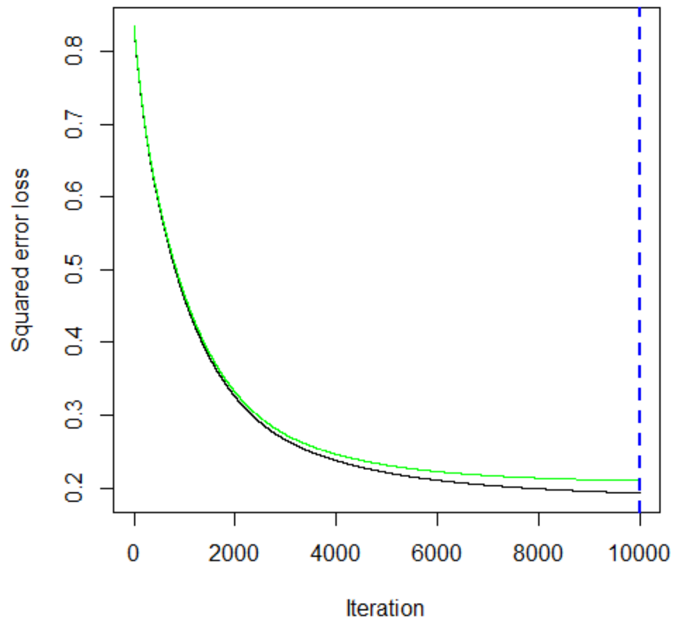


GRADIENT BOOSTING MACHINE

- The first gradient boosting model to predict ln(expenses) is trained based on the below list of parameters:

Parameter	Value
Number of trees	10,000
Depth of each tree	1
Learning rate (Shrinkage)	0.001
CV (cross-validation) folds	5

- All variables had non-zero influence.
- The algorithm used 9,996 trees with an RMSE of 0.46 on the training dataset.



GRADIENT BOOSTING MACHINE

- A grid search is a better alternative than manually modifying hyperparameters one at a time to determine the best combination.
- The search will be conducted across 81 models, based on the below hyperparameter grid:

Parameter	Varying Values
Learning rate (Shrinkage)	0.01, 0.1 and 0.3
Depth of each tree	1, 3 and 5
Minimum number of observations allowed in the trees' terminal nodes	5, 10 and 15
Subsampling fraction	0.65, 0.8 and 1

- Instead of performing 5-fold CV, 75% of the training observations are used and performance is evaluated on the remaining 25% to speed up the tuning process and reduce computation time.

GRADIENT BOOSTING MACHINE

Model	Learning Rate	Depth of each tree	Min. # of observations in the terminal nodes	Subsampling fraction	Optimal Number of trees	Minimum RMSE
1	0.1	5	5	1	45	0.37572
2	0.1	5	10	1	46	0.37621
3	0.01	5	5	1	453	0.37713
4	0.1	5	15	1	53	0.37718
5	0.01	5	10	1	467	0.37719
6	0.1	5	15	0.65	48	0.37722
7	0.3	5	15	1	15	0.37729
8	0.1	5	15	0.8	43	0.37734
9	0.01	5	15	1	528	0.37776
10	0.3	5	5	1	12	0.37850

- The training fraction is now set to 1, and the RMSE of the final training dataset records 0.34.

GRADIENT BOOSTING MACHINE

- The final model is applied to predict on the testing dataset.

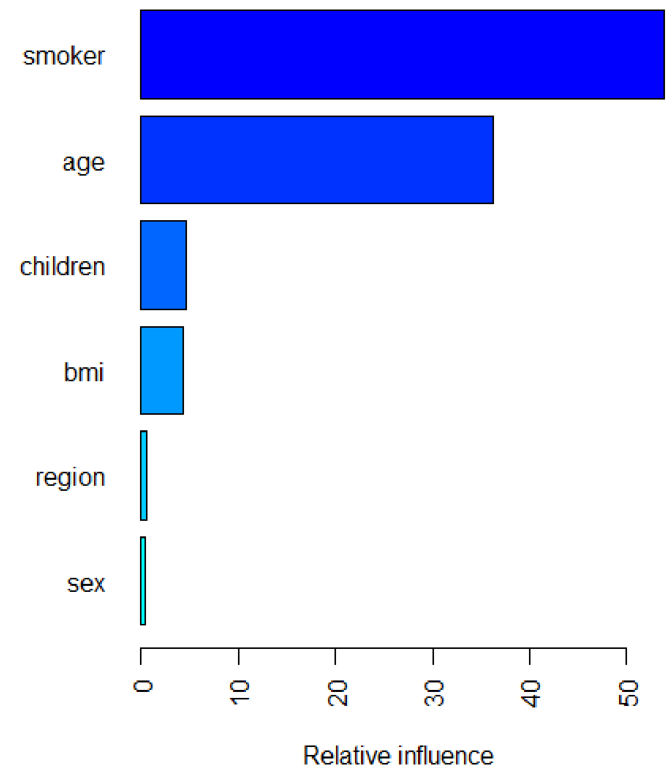
RMSE	MAE	MAPE
0.378	0.22	2.4%

- The recorded accuracy is 97.60%.

GRADIENT BOOSTING MACHINE

Variable Importance of the GBM model

Variables	Relative Influence
Smoker	53.89%
Age	36.21%
Children	4.59%
BMI	4.29%
Region	0.62%
Sex	0.40%

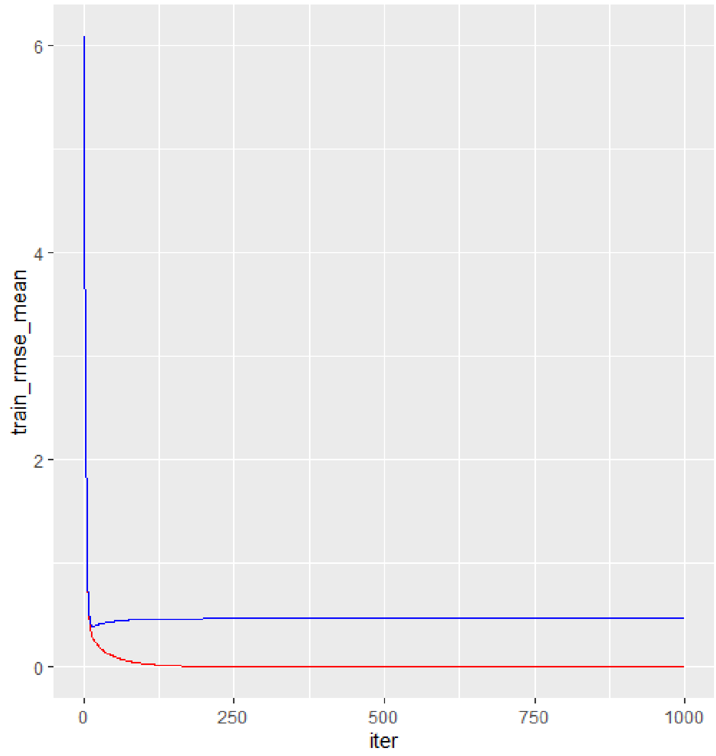


XGBOOST

- The first XGBoost model to predict ln(expenses) is trained based on the below list of parameters:

Parameter	Value
Number of trees	1,000
Depth of each tree	6
Learning rate (Shrinkage)	0.30
Minimum node size	1
Subsampling fraction	100%
CV (cross-validation) folds	5

- The algorithm used 14 trees only with an RMSE of 0.39 on the training dataset.



XGBOOST

- A grid search is a better alternative than manually modifying hyperparameters one at a time to determine the best combination.
- Similar to GBM, a grid search will be conducted across 576 models, based on the below hyperparameter grid:

Parameter	Varying Values
Learning rate (Shrinkage)	0.01, 0.05, 0.1 and 0.3
Depth of each tree	1, 3, 5 and 7
Minimum number of observations allowed in the trees' terminal nodes	1, 3, 5 and 7
Subsampling fraction	0.65, 0.8 and 1
Percent of columns to sample from for each tree	0.65, 0.8 and 1

MODELING RESULTS

XGBOOST

Model	Learning Rate	Depth of each tree	Min. # of observations in the terminal nodes	Subsampling fraction	Percent of columns to sample from for each tree	Optimal Number of trees	Minimum RMSE
1	0.1	3	7	1	0.8	69	0.37208
2	0.1	3	3	1	0.9	65	0.37241
3	0.1	3	5	1	0.9	63	0.37259
4	0.01	3	7	1	0.8	668	0.37264
5	0.1	3	5	1	0.8	69	0.37272
6	0.01	3	5	1	0.8	677	0.37280
7	0.1	3	1	1	0.9	69	0.37286
8	0.01	3	3	1	0.8	660	0.37302
9	0.05	3	7	1	0.9	129	0.37321
10	0.1	3	7	1	0.9	61	0.37327

The RMSE of the top trained model records 0.37208, less than the first XGBoost model.

XGBOOST

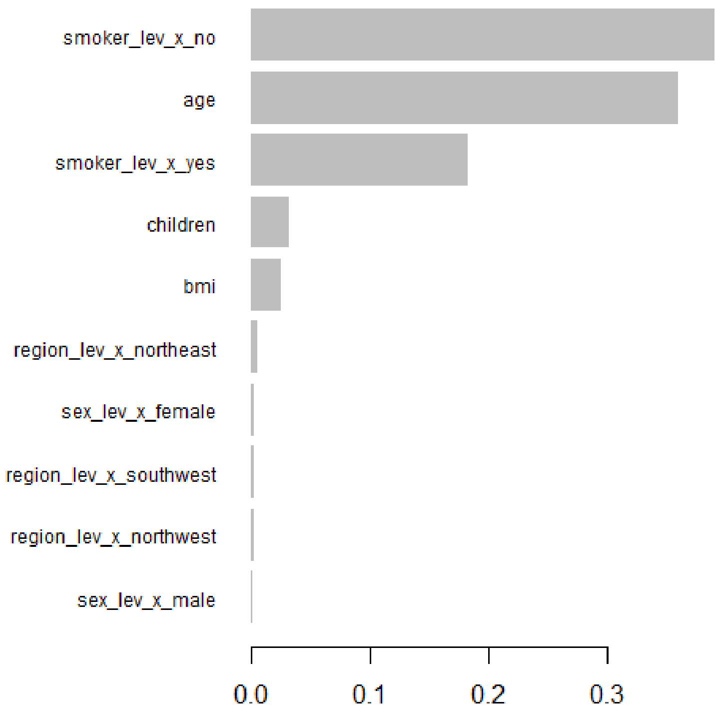
- ❑ The final model is applied to predict on the testing dataset.

RMSE	MAE	MAPE
0.376	0.20	2.26%

- ❑ The recorded accuracy is 97.74%.

XGBOOST

- ❑ XGBoost algorithm includes a variable importance plotting feature as well.
- ❑ Smoker and Age are the most important variables, followed by children and BMI. Region and Sex have negligible importance.



MULTIPLE LINEAR REGRESSION

- The first MLR model includes all the independent variables, and the second MLR includes the important variables.

$$ln(expenses) = 5.60 + 0.0348 \times Age - 0.0689 \times Sex + 0.0143 \times BMI + 0.103 \times Children + 1.52 \times Smoker - 0.0579 \times Region$$

$$ln(expenses) = 5.41 + 0.0349 \times Age + 0.0125 \times BMI + 0.103 \times Children + 1.51 \times Smoker$$

lmmod1.ln	Estimated coefficient	t-test value	p-value
(Intercept)	5.60	55.40	1.59e-296
Age	0.0348	33.70	1.74e-163
Sex	-0.0689	-2.37	1.79e-2
BMI	0.0143	6.04	2.20e-9
Children	0.103	8.60	3.37e-17
Smoker	1.52	42.10	2.30e-217
Region	-0.0579	-4.45	9.80e-6

lmmod2.ln	Estimated coefficient	t-test value	p-value
(Intercept)	5.41	58.80	1.53e-315
Age	0.0349	33.50	3.17e-162
BMI	0.0125	5.30	1.46e-7
Children	0.103	8.43	1.30e-16
Smoker	1.51	41.40	1.32e-213

- In both models, all variables are shown to be significant with p-values less than 0.05.

MULTIPLE LINEAR REGRESSION

- The models are applied to predict on the testing dataset.

Model	R-squared	Adjusted R-squared	MAPE	AIC	BIC
MLR with all variables	0.77	0.77	3.12%	1,133.21	1,171.94
MLR with important variables	0.76	0.76	3.14%	1,154.50	1,183.55

- The accuracy of the first model is 96.88%, and that of the second model is 96.86%.
- The accuracy of the first model is very close but slightly better than that of the second model (The first model includes all the independent variables).

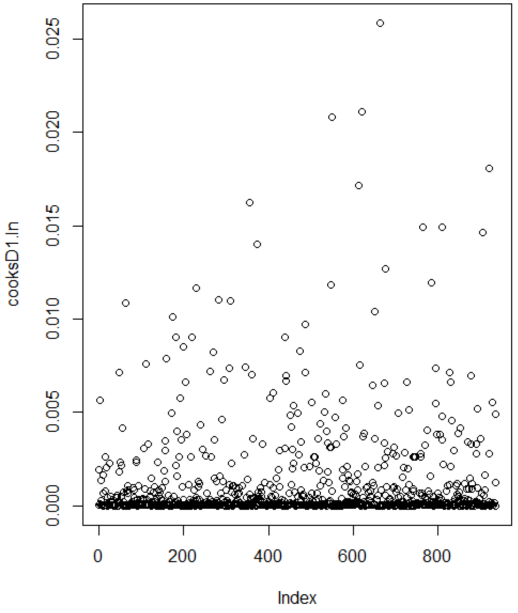
MULTIPLE LINEAR REGRESSION

Cook’s Distance

- Influential observations in a dataset can affect a model’s performance.
- Cook’s distance is applied to both linear regression models to determine the influential data points.
- It summarizes how much a regression model changes when the i^{th} observation is deleted.
- Generally, any point with a Cook's Distance greater than 4/n is regarded as an outlier.

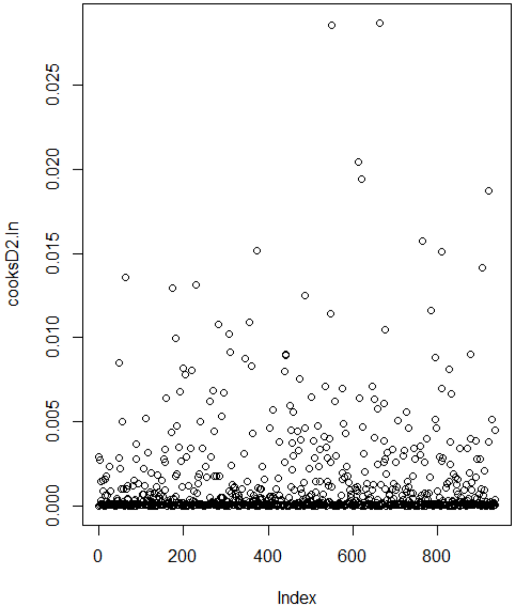
MULTIPLE LINEAR REGRESSION

Cook's Distance of the first linear model



69 influential observations

Cook's Distance of the second linear model



66 influential observations

MULTIPLE LINEAR REGRESSION

- The influential observations of each model are removed.

$$\ln(expenses) = 5.40 + 0.0389 \times Age - 0.0988 \times Sex + 0.00983 \times BMI + 0.108 \times Children + 1.60 \times Smoker - 0.0293 \times Region$$

$$\ln(expenses) = 5.32 + 0.0384 \times Age + 0.00747 \times BMI + 0.104 \times Children + 1.56 \times Smoker$$

lmmod1.ln (cooks)	Estimated coefficient	t-test value	p-value
(Intercept)	5.40	66.30	0.00
Age	0.0389	47.10	4.32e-243
Sex	-0.0988	-4.36	1.46e-5
BMI	0.00983	5.10	4.25e-7
Children	0.108	11.50	1.46e-28
Smoker	1.60	55.80	2.99e-291
Region	-0.0293	-2.85	4.44e-3

lmmod2.ln (cooks)	Estimated coefficient	t-test value	p-value
(Intercept)	5.32	67.10	0.00
Age	0.0384	41.80	1.48e-211
BMI	0.00747	3.55	3.98e-4
Children	0.104	9.81	1.26e-21
Smoker	1.56	50.90	2.36e-265

- In both models, all variables are shown to be significant with p-values less than 0.05.

MULTIPLE LINEAR REGRESSION

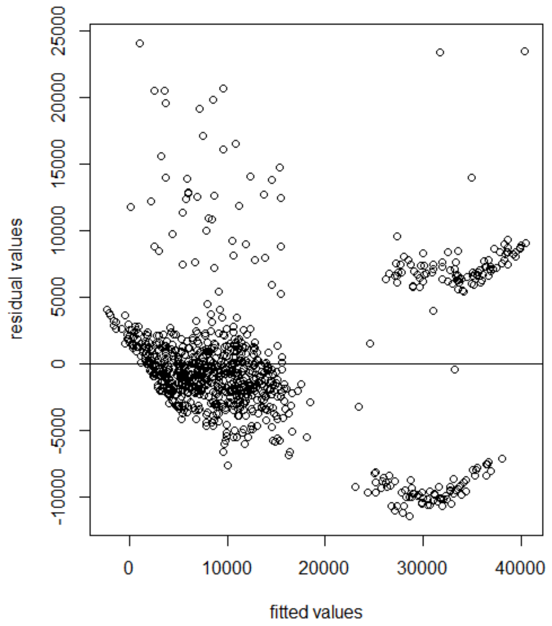
- The models are applied to predict on the testing dataset.

Model	R-squared	Adjusted R-squared	MAPE	AIC	BIC
MLR with all variables	0.77	0.77	3.12%	1,133.21	1,171.94
MLR with important variables	0.76	0.76	3.14%	1,154.50	1,183.55
MLR(Cooks) with all variables	**0.86**	**0.86**	**2.40%**	**592.21**	**630.52**
MLR(Cooks) with important variables	**0.84**	**0.84**	**2.29%**	**758.19**	**786.94**

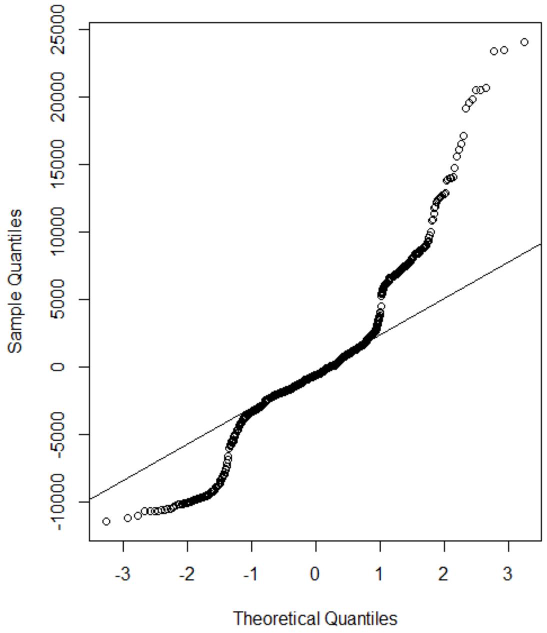
- The R-squared has improved from 77% to 86% for the first model, and from 76% to 84% for the second model.
- The second model without influential observations records the best accuracy of 97.71%.

MULTIPLE LINEAR REGRESSION

Residual Vs. Fitted Values Plot



Q-Q Plot

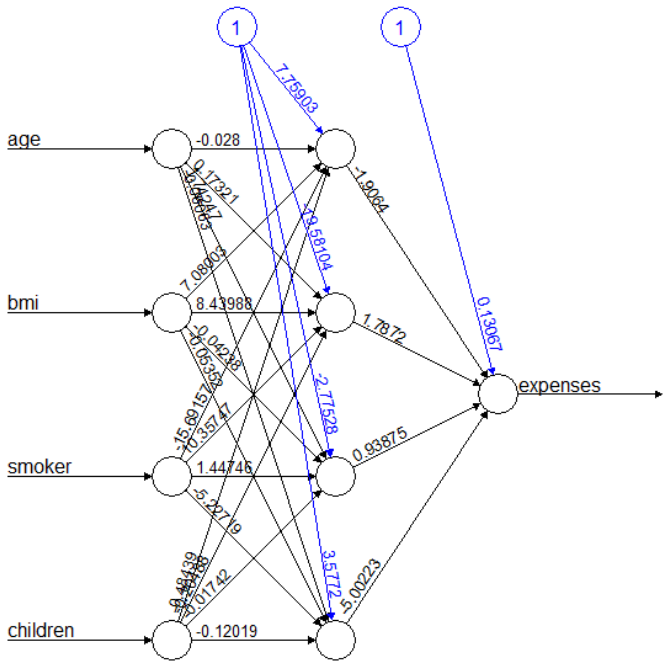


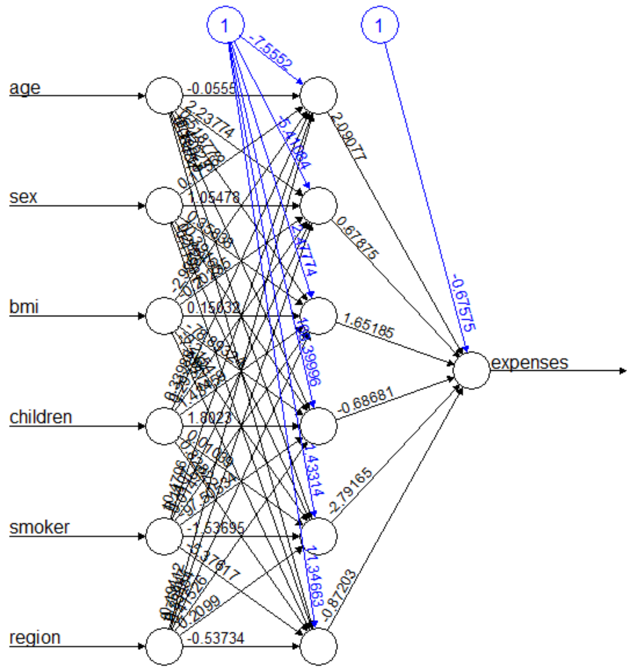
BP value	P-value
30.13	4.613e-6

The p-value of the BP test is less than 0.05, indicating heteroscedasticity is present.

FEEDFORWARD NEURAL NETWORK

Expenses





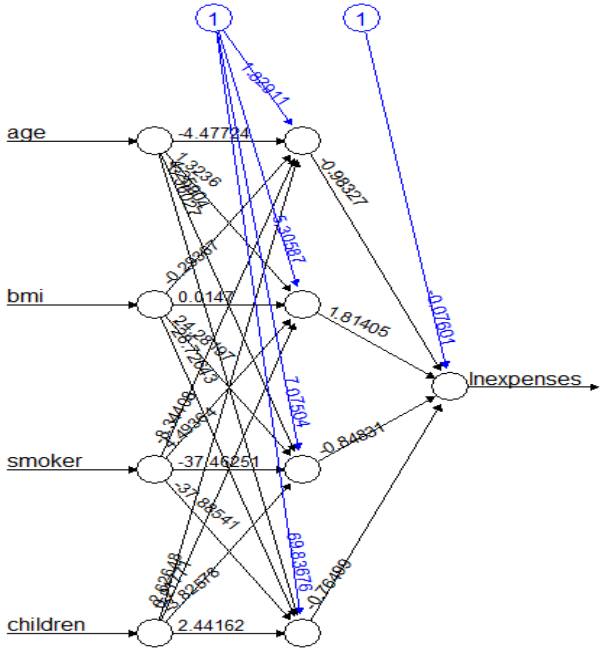
FNN 1: The input layer has 4 elements (the important variables of the dataset) and 4 neurons in the hidden layer are assumed.

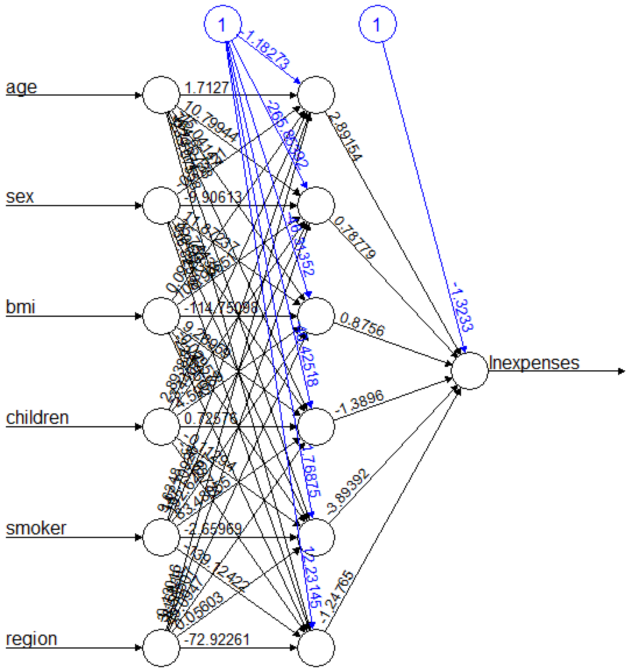
FNN 2: The input layer has 6 elements (all independent variables of the dataset) and 6 neurons in the hidden layer are assumed.

MODELING RESULTS

FEEDFORWARD NEURAL NETWORK

Ln(Expenses)



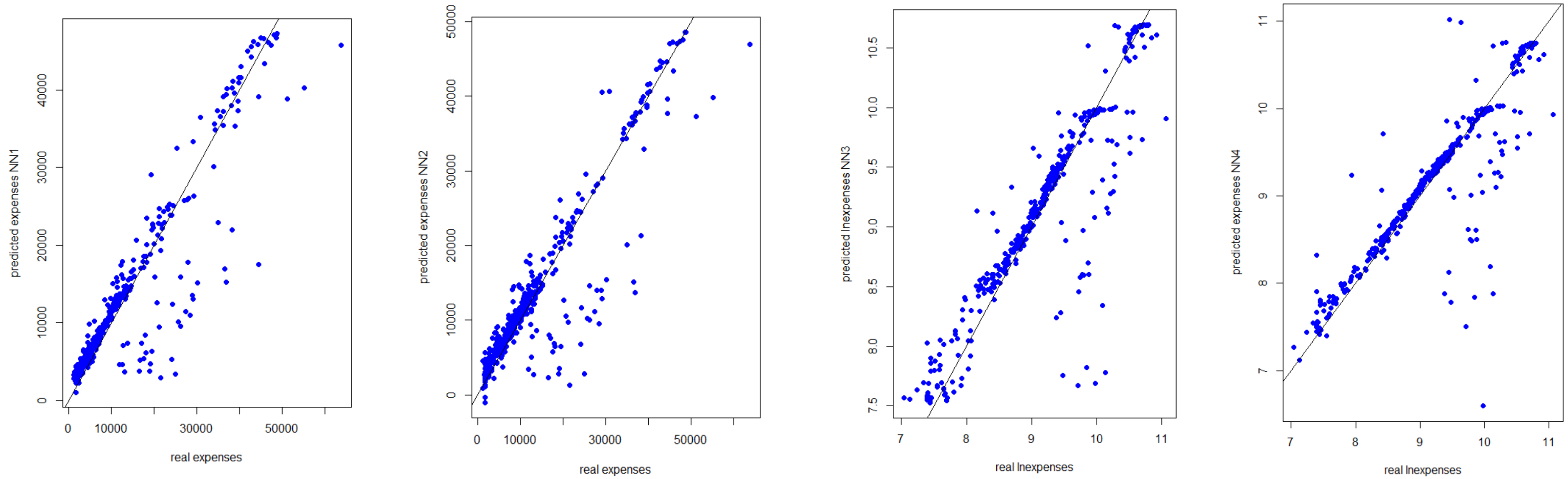


FNN 3: The input layer has 4 elements (the important variables of the dataset) and 4 neurons in the hidden layer are assumed.

FNN 4: The input layer has 6 elements (all independent variables of the dataset) and 6 neurons in the hidden layer are assumed.

MODELING RESULTS

FEEDFORWARD NEURAL NETWORK



Model	FNN 1	FNN 2	FNN 3	FNN 4
RMSE	4,793.20	4,809.31	0.39	0.42
MAPE	29.13%	28.44%	2.29%	2.11%
Accuracy	70.87%	71.56%	97.71%	97.89%

AGENDA

Introduction

Literature Review

Methodology

Modeling Results

Conclusion

Future Work

Questions

CONCLUSION

- Predicting healthcare expenses is a topic that is continuously researched.
- Several machine models are applied to determine the best one for predicting healthcare expenses.

Model	Accuracy	
	Expenses	Ln(Expenses)
Regression Decision Tree	50.53%	97.30%
Gradient Boosting Machine	65.00%	97.60%
XGBoost	67.87%	97.74%
Multiple Linear Regression	60.25%	97.71%
Feedforward Neural Network	71.56%	97.89%

- Feedforward Neural Network outperformed the remaining models in both cases.

AGENDA

- Introduction
- Literature Review
- Methodology
- Modeling Results
- Conclusion
- **Future Work**
- Questions

FUTURE WORK

To further improve the study's results:

- ❑ The study can aim to include a much larger dataset to minimize problems in the distribution of expenses.
- ❑ The dataset can include more variables that might affect expenses.
- ❑ Credibility will be applied to determine the minimum sample size required.
- ❑ More complex and sophisticated models can be applied on the dataset.
- ❑ The expertise of health insurance policymakers can be incorporated during the application of the models.

AGENDA

- Introduction
- Literature Review
- Methodology
- Modeling Results
- Conclusion
- Future Work
- **Questions**

- Associate Professor at Notre Dame University Louaize and acts as an Academic Advisor for graduate and undergraduate programs in Actuarial Sciences.
- Possesses over ten years of experience in Modeling, Estimation, Prediction, Analysis, Econometrics, and Computational Statistics.
- Research interests and teaching topics include Nonparametric Estimation, Statistical Learning, Statistical Inference, Computational Statistics, Optimization, Machine Learning/deep learning, and Modeling in various fields such as engineering, finance, biology/health/medicine, and actuarial science.

ABOUT ME

Re-Mi
Hage

Notre Dame
University Louaize

ABOUT ME

Liana Barsoumian

Notre Dame University Louaize

- Senior Underwriter at SNA S.A.L. Lebanon with almost 3 years of experience in the insurance industry.
- Holds a Bachelor and Master's degree in Actuarial Sciences from Notre Dame University – Louaize.
- Interests include Machine Learning, Statistical Learning, Modeling, Predictive Analytics and Analysis, especially in the field of Life and Health Insurance.

EAA e-Conference on
Data Science & Data Ethics

14 May 2024

Thank you very much for your attention

Contact

Liana Barsoumian – Re-Mi Hage

Notre Dame University Louaize

lobarsoumian@ndu.edu.lb

rhage@ndu.edu.lb