DATA AND
ARTIFICIAL INTELLIGENCE

Data Science Capstone

simplilearn

DATA AND
ARTIFICIAL INTELLIGENCE

Healthcare Insurance Analysis

simplilearn

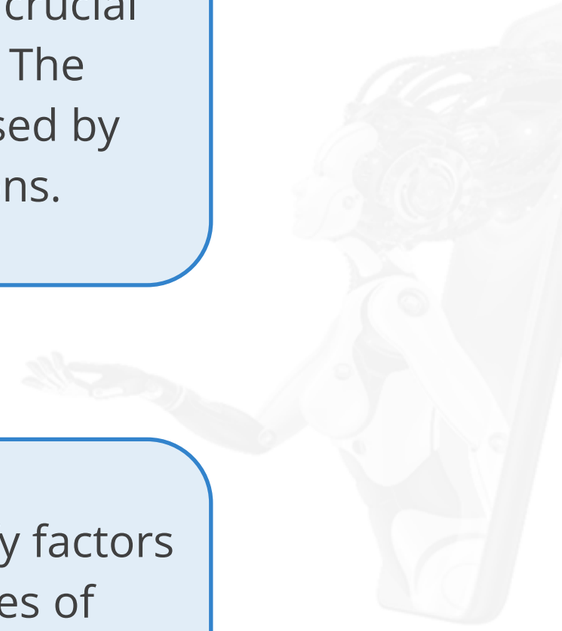
Business Scenario

Problem statement:

A significant public health concern is the rising cost of healthcare. Therefore, it's crucial to be able to predict future costs and gain a solid understanding of their causes. The insurance industry must also take this analysis seriously. This analysis may be used by healthcare insurance providers to make a variety of strategic and tactical decisions.

Objective:

The objective of this project is to predict patients' healthcare costs and to identify factors contributing to this prediction. It will also be useful to learn the interdependencies of different factors and comprehend the significance of various tools at various stages of the healthcare cost prediction process.



Dataset Snapshot

Hospitalization details.xlsx

1	Customer ID	year	month	date	children	charges	Hospital tier	City tier	State ID
2	Id2335	1992	Jul	9	0	563.84	tier - 2	tier - 3	R1013
3	Id2334	1992	Nov	30	0	570.62	tier - 2	tier - 1	R1013
4	Id2333	1993	Jun	30	0	600	tier - 2	tier - 1	R1013
5	Id2332	1992	Sep	13	0	604.54	tier - 3	tier - 3	R1013
6	Id2331	1998	Jul	27	0	637.26	tier - 3	tier - 3	R1013
7	Id2330	2001	Nov	20	0	646.14	tier - 3	tier - 3	R1012
8	Id2329	1993	Jun	1	0	650	tier - 3	tier - 3	R1013
9	Id2328	1995	Jul	4	0	650	tier - 3	tier - 3	R1013
10	Id2327	2002	Nov	29	0	668	tier - 3	tier - 2	R1012
11	Id2326	1997	Nov	9	0	670	tier - 3	tier - 3	R1013
12	Id2325	2001	Sep	12	0	687.54	tier - 3	tier - 2	R1013
13	Id2324	1999	Dec	26	0	700	?	tier - 3	R1013
14	Id2323	1999	Dec	14	0	722.99	tier - 3	tier - 1	R1013
15	Id2322	2002	?	19	0	750	tier - 3	tier - 1	R1012
16	Id2321	1993	Aug	9	0	760	tier - 3	tier - 1	R1013
17	Id2320	1996	Oct	22	0	760	tier - 3	tier - 3	R1013
18	Id2319	1993	Jun	28	0	770	tier - 3	tier - 3	R1013
19	Id2318	1996	?	18	0	770.38	tier - 3	?	R1012
20	Id2317	1995	Dec	7	0	773.54	tier - 3	tier - 2	R1013

Dataset Description

Hospitalization details.xlsx

Variables	Description
Customer ID	Unique identification for beneficiary(primary)
year	Year of birth
month	Month of birth
date	Date of birth
children	No. of children as dependents
charges	Hospitalization cost
Hospital tier	Level of hospital, tier 1 being the best
City tier	Level of city per government document, tier 1 referring to the most developed
State ID	ID of the state

Dataset Snapshot

Medical Examinations.xlsx

1	Customer ID	BMI	HBA1C	Heart Issues	Any Transplants	Cancer history	NumberOfMajorSurgeries	smoker
2	Id1	47.41	7.47	No	No	No	No major surgery	yes
3	Id2	30.36	5.77	No	No	No	No major surgery	yes
4	Id3	34.485	11.87	yes	No	No	2	yes
5	Id4	38.095	6.05	No	No	No	No major surgery	yes
6	Id5	35.53	5.45	No	No	No	No major surgery	yes
7	Id6	32.8	6.59	No	No	No	No major surgery	yes
8	Id7	36.4	6.07	No	No	No	No major surgery	yes
9	Id8	36.96	7.93	No	No	No	3	yes
10	Id9	41.14	9.58	yes	No	Yes	1	yes
11	Id10	38.06	10.79	No	No	No	No major surgery	yes
12	Id11	37.7	5.96	yes	No	No	2	yes
13	Id12	42.13	11.9	No	No	No	No major surgery	yes
14	Id13	40.92	8.41	No	No	No	No major surgery	yes
15	Id14	40.565	7.02	No	No	No	No major surgery	yes
16	Id15	36.385	7.59	yes	No	No	2	yes
17	Id16	39.9	11.32	No	No	No	No major surgery	yes
18	Id17	33.8	7.67	No	No	No	3	yes
19	Id18	36.765	7.29	yes	No	Yes	1	yes
20	Id19	36.955	4.72	yes	No	No	1	yes
21	Id20	42.9	11.41	No	No	No	No major surgery	yes
22	Id21	36.3	11.5	yes	No	No	2	yes

Dataset Description

Medical Examinations.xlsx

Variables	Description
Customer ID	Unique identification for beneficiary(primary)
BMI	Shows the body mass index of the individual (BMI measures body fat based on height and weight)
HBA1C	Shows the HBA1C report (HBA1C measures the amount of sugar in the blood (glucose), where HBA1C greater than 6.5 is considered diabetic
Heart Issues	Shows if a patient has heart-related issues
Any Transplants	Shows if a patient has any transplants in their body
Cancer history	Shows if a patient has any history of cancer in the family
NumberOfMajorSurgeries	Displays the number of major surgeries a patient has gone through
smoker	Indicates if a patient smokes cigarettes

Dataset Snapshot

Names.xlsx

	Customer ID	name
2	Id1	Hawks, Ms. Kelly
3	Id2	Lehner, Mr. Matthew D
4	Id3	Lu, Mr. Phil
5	Id4	Osborne, Ms. Kelsey
6	Id5	Kadala, Ms. Kristyn
7	Id6	Baker, Mr. Russell B.
8	Id7	Macpherson, Mr. Scott
9	Id8	Hallman, Mr. Stephen
10	Id9	Moran, Mr. Patrick R.
11	Id10	Benner, Ms. Brooke N.
12	Id11	Fierro Vargas, Ms. Paola Andrea
13	Id12	Franz, Mr. David
14	Id13	Foster, Mr. Wade
15	Id14	Tenorio, Mr. Franklin
16	Id15	Rios, Ms. Leilani M.
17	Id16	Viau-Dupuis, Mr. Philippe

Dataset Description

Names.xlsx

Variables	Description
Customer ID	Unique identification for beneficiary(primary)
name	Name of the beneficiary(primary)

Project Task: Week 1

Data science/data analysis

1. Collate the files so that all the information is in one place
2. Check for missing values in the dataset
3. Find the percentage of rows that have trivial value (for example, *?),* and delete such rows if they do not contain significant information
4. Use the necessary transformation methods to deal with the nominal and ordinal categorical variables in the dataset
5. The dataset has *State ID,* which has around 16 states. All states are not represented in equal proportions in the data. Creating dummy variables for all regions may also result in too many insignificant predictors. Nevertheless, only R1011, R1012, and R1013 are worth investigating further. Create a suitable strategy to create dummy variables with these restraints.
6. The variable *NumberOfMajorSurgeries* also appears to have string values. Apply a suitable method to clean up this variable.

Note: Use Excel as well as Python to complete the tasks

Project Task: Week 1

Data science/data analysis

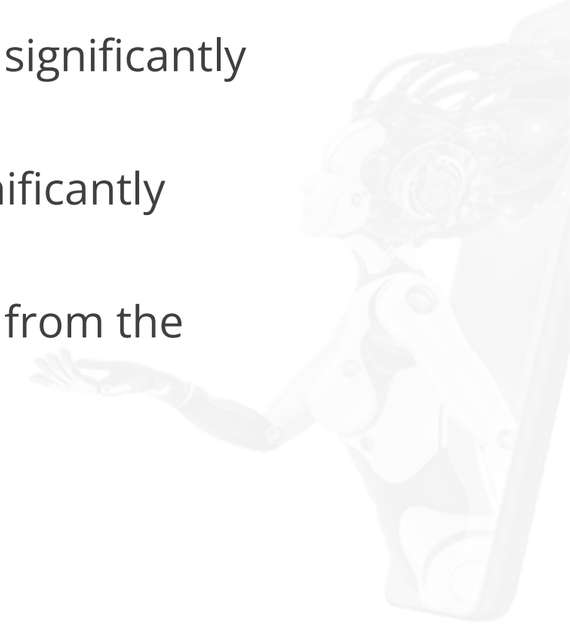
7. *Age* appears to be a significant factor in this analysis. Calculate the patients' ages based on their dates of birth.
8. The gender of the patient may be an important factor in determining the cost of hospitalization. The salutations in a beneficiary's name can be used to determine their gender. Make a new field for the beneficiary's gender.
9. You should also visualize the distribution of costs using a histogram, box and whisker plot, and swarm plot.
10. State how the distribution is different across gender and tiers of hospitals
11. Create a radar chart to showcase the median hospitalization cost for each tier of hospitals
12. Create a frequency table and a stacked bar chart to visualize the count of people in the different tiers of cities and hospitals

Note: Use Excel as well as Python to complete the tasks

Project Task: Week 1

Data science/data analysis

13. Test the following null hypotheses:
    a. The average hospitalization costs for the three types of hospitals are not significantly different
    b. The average hospitalization costs for the three types of cities are not significantly different
    c. The average hospitalization cost for smokers is not significantly different from the average cost for nonsmokers
    d. Smoking and heart issues are independent



Note: Use Excel as well as Python to complete the tasks

Project Task: Week 2

Machine Learning

1. Examine the correlation between predictors to identify highly correlated predictors. Use a heatmap to visualize this.
2. Develop and evaluate the final model using regression with a stochastic gradient descent optimizer. Also, ensure that you apply all the following suggestions:

Note:

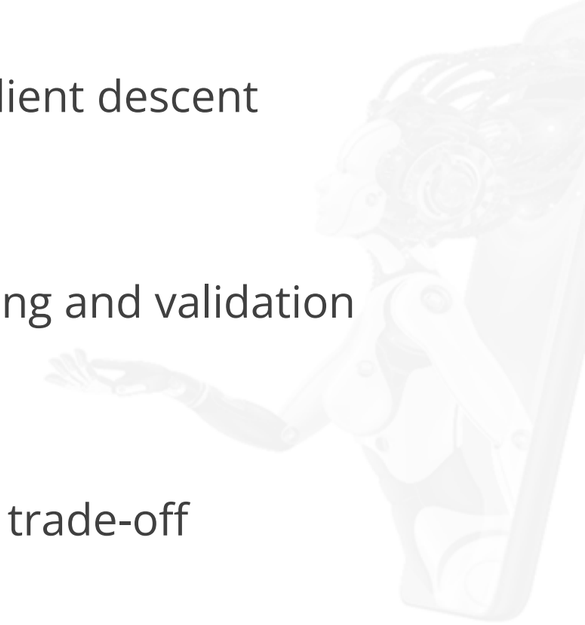
- Perform the stratified 5-fold cross-validation technique for model building and validation
- Use standardization and hyperparameter tuning effectively
- Use sklearn-pipelines
- Use appropriate regularization techniques to address the bias-variance trade-off

a. Create five folds in the data, and introduce a variable to identify the folds

b. For each fold, run a *for* loop and ensure that 80 percent of the data is used to train the model and the remaining 20 percent is used to validate it in each iteration

c. Develop five distinct models and five distinct validation scores (root mean squared error values)

d. Determine the variable importance scores, and identify the redundant variables



Project Task: Week 2

Machine Learning

3. Use random forest and extreme gradient boosting for cost prediction, share your cross-validation results, and calculate the variable importance scores
4. Case scenario:

   Estimate the cost of hospitalization for Christopher, Ms. Jayna (her date of birth is 12/28/1988, height is 170 cm, and weight is 85 kgs). She lives in a tier-1 city and her state's *State ID* is R1011. She lives with her partner and two children. She was found to be nondiabetic (HbA1c = 5.8). She smokes but is otherwise healthy. She has had no transplants or major surgeries. Her father died of lung cancer. Hospitalization costs will be estimated using tier-1 hospitals.
5. Find the predicted hospitalization cost using all five models. The predicted value should be the mean of the five models' predicted values.

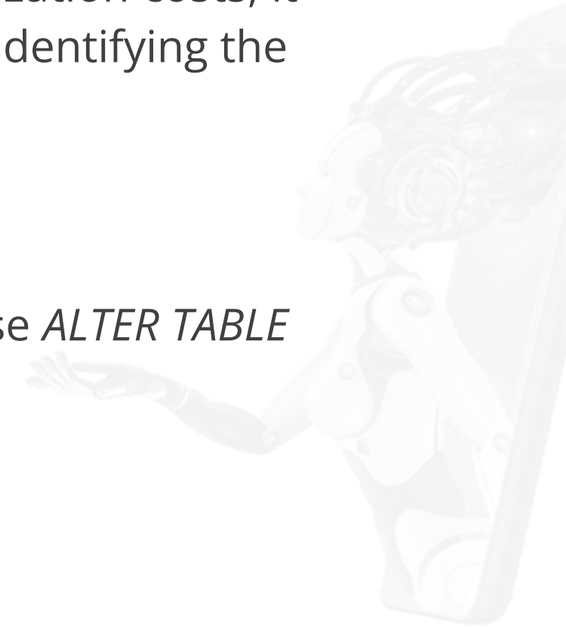
Project Task: Week 2

SQL

1. To gain a comprehensive understanding of the factors influencing hospitalization costs, it is necessary to combine the tables provided. Merge the two tables by first identifying the columns in the data tables that will help you in merging.

   a. In both tables, add a *Primary Key* constraint for these columns

Hint: You can remove duplicates and null values from the column and then use *ALTER TABLE* to add a *Primary Key* constraint.



Project Task: Week 2

SQL

2. Retrieve information about people who are diabetic and have heart problems with their average age, the average number of dependent children, average BMI, and average hospitalization costs
3. Find the average hospitalization cost for each hospital tier and each city level
4. Determine the number of people who have had major surgery with a history of cancer
5. Determine the number of tier-1 hospitals in each state

Project Task: Week 2

Tableau

1. Create a dashboard in Tableau by selecting the appropriate chart types and business metrics

Note: Put more emphasis on data storytelling

Thank You