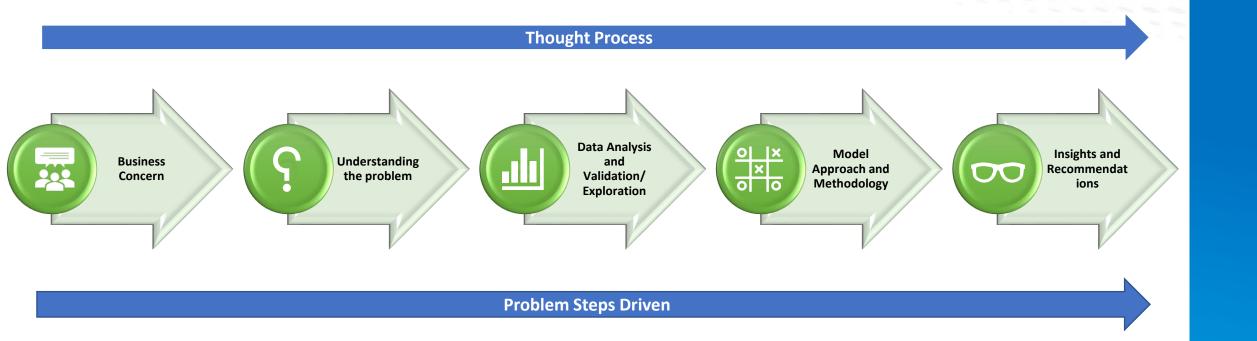
greatlearning
Power Ahead

Capstone Presentation



Roadmap of the Presentation





Business Problem Understanding

Business Problem → Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance, then it will become a tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced.

Objective → To build a model, using data that provide the optimum insurance cost for an individual.

Tools used → Python & Tableau









Data Preparation

Step 1: Data Extraction

 Extracting data with 25000 entries from CSV file using Pandas library.



Step 2: Data Cleaning

- Missing values analysis and Treatment.
- Handling Outliers
- Change Columns names to make reading easier.



Step 3: Feature Engineering

 Transforming raw data into features that can be used in supervised learning



Brief Summary of Data

- 25000 Rows & 24 Columns.
- 8 Object, 2 Float & 14 Int data types.
- Missing values: bmi 990,
 Year_last_admitted 11,881

applicant_id	1	-0.017	-0.0083	0.0039	-0.0032	0.0012	0.00079	-0.0056	0.0045	-0.0014	0.0041	-0.0038	0.0064	0.011	-0.0014
years_of_insurance_with_us	-0.017	1	0.019	0.015	0.0041	-0.0016	0.0038	-0.0039	-0.011	-0.0016	-0.013	-0.0061	-0.00031	-0.0031	0.0014
regular_checkup_last_year	-0.0083	0.019	1	0.0095	-0.0063	-0.0028	0.0086	-0.0035	-0.01	0.016	-0.0086	-0.14	-0.011	0.003	-0.18
adventure_sports	0.0039	0.015	0.0095	1	0.0098	-0.0014	-0.0025	0.00014	0.0026	-0.0046	-0.0023	0.074	-0.049	0.0037	0.074
visited_doctor_last_1_year	-0.0032	0.0041	-0.0063	0.0098	1	-0.16	-0.0006	0.00031	0.0094	0.0067	0.00052	0.012	-0.012	-0.045	0.0083
daily_avg_steps	0.0012	-0.0016	-0.0028	-0.0014	-0.16	1	-0.00064	0.004	-0.0031	-0.00026	-0.0057	-0.0045	0.0073	0.047	-0.005
age	0.00079	0.0038	0.0086	-0.0025	-0.0006	-0.00064	1	-0.0046	0.0025	-0.011	-0.015	0.00028	-0.005	-0.011	0.004
heart_decs_history	-0.0056	-0.0039	-0.0035	0.00014	0.00031	0.004	-0.0046	1	0.11	-0.0043	0.041	-0.0025	0.0041	0.00034	0.0014
other_major_decs_history	0.0045	-0.011	-0.01	0.0026	0.0094	-0.0031	0.0025	0.11	1	-0.00074	0.17	-0.0027	0.0016	0.0046	-0.001
avg_glucose_level	-0.0014	-0.0016	0.016	-0.0046	0.0067	-0.00026	-0.011	-0.0043	-0.00074	1	-0.019	-0.0038	2.7e-05	0.0021	-0.003
bmi	0.0041	-0.013	-0.0086	-0.0023	0.00052	-0.0057	-0.015	0.041	0.17	-0.019	1	-0.0078	0.018	-0.0032	-0.008
weight	-0.0038	-0.0061	-0.14	0.074	0.012	-0.0045	0.00028	-0.0025	-0.0027	-0.0038	-0.0078	1	-0.37	-0.0068	0.97
weight_change_in_last_one_year	0.0064	-0.00031	-0.011	-0.049	-0.012	0.0073	-0.005	0.0041	0.0016	2.7e-05	0.018	-0.37	1	0.013	-0.34
fat_percentage	0.011	-0.0031	0.003	0.0037	-0.045	0.047	-0.011	0.00034	0.0046	0.0021	-0.0032	-0.0068	0.013	1	-0.008
insurance_cost	-0.0014	0.0014	-0.18	0.074	0.0083	-0.0052	0.0043	0.0014	-0.0013	-0.0039	-0.0082	0.97	-0.34	-0.0081	1
	applicant_id	ears_of_insurance_with_us	regular_checkup_last_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	other_major_decs_history	avg_glucose_level	bmi	weight	_change_in_last_one_year	fat_percentage	insurance_cost



Model Approached

Linear regression model, ANN, Decision Tree and Random Forest model is created using Scikit and stats model packages. Output from Stats Model and using comparison tables is as follows.

Values for RMSE models is extreme high, we cannot consider, we will try to perform model tuning see the results and interpret the best model to optimize the insurance cost per individual. Values for R-square models are good score. We can see and compare their scores to interpret the business model. For now, lets do the grid search for each model and find the best parameters to identify the accurate model.

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	3375.862926	3339.684420	0.944401	0.945956
Decision Tree Regressor	0.000000	4356.047598	1.000000	0.908056
Random Forest Regressor	1166.608857	3092.671229	0.993360	0.953655
ANN Regressor	3009.981169	3093.412815	0.955800	0.953632



Insights from Analysis

- From the observations, we can say that we can add the variable or parameter named as premium can be added as adding up there average expenses or their income from the occupation depends on location as well.
- As there are more than 50 % missing values in year last admitted parameter, it will highly impact the accuracy score, we dropped the columns.
- As per the model approach, Random Forest has good accuracy score and R value.



Recommendations

- They must introduce the Skills upgrading scheme which may help individual to organize their day-to- day activities in a structured way. Such as it should include managerial, administrative, technical and social skills.
- Managerial skills to manage the entire program
- Administrative skills to manage finances and the funds
- Technical skills to understand the complexities of health insurance
- Social skills to understand the community's needs To optimize the best insurance cost per individual they must execute the 15 to 30 days routine check-up by organizing medical camps and awareness program.
- There must be a network of health care providers (public or private). Without this, it is not wise to talk about health insurance.
- The people must have the capacity to pay the premium. Unless there will be no takes for health insurance.
- There are many more recommendations which needs to be fulfilled but at least you should focus on executing the three recommendations to introduce the healthcare insurance programme.



Thankyou

