# HEALTH INSURANCE COST, CAPSTONE PROJECT

*Presenter:* ***Priyamvada***

Dated: 03/11/2023 (dd/mm/yyyy)

# Business Problem

- Insurance companies can run into losses if they don’t optimize their health insurance cost for the customers. This is because customers have an individual healthcare cost depending on their lifestyle choices and pre-existing health risks. To minimize its business risk, the insurance company needs a reliable way to predict estimate insurance cost for all individuals in its database.

- To achieve this, the company wants to build a predictive model that will calculate the insurance cost for each customer using different lifestyle and health parameters.

# Data Dictionary

| Variable | Business Definition |
|---|---|
| applicant_id | Applicant unique ID |
| years_of_insurance_with_us | Since how many years customer is taking policy from the same company only |
| regular_checkup_lasy_year | Number of times customers has done the regular health check up in last one year |
| adventure_sports | Customer is involved with adventure sports like climbing, diving etc. |
| Occupation | Occupation of the customer |
| visited_doctor_last_1_year | Number of times customer has visited doctor in last one year |
| cholesterol_level | Cholesterol level of the customers while applying for insurance |
| daily_avg_steps | Average daily steps walked by customers |
| age | Age of the customer |
| heart_decs_history | Any past heart diseases |
| other_major_decs_history | Any past major diseases apart from heart like any operation |
| Gender | Gender of the customer |
| avg_glucose_level | Average glucose level of the customer while applying the insurance |
| bmi | BMI of the customer while applying the insurance |
| smoking_status | Smoking status of the customer |
| Year_last_admitted | When customer have been admitted in the hospital last time |
| Location | Location of the hospital |
| weight | Weight of the customer |
| covered_by_any_other_company | Customer is covered from any other insurance company |
| Alcohol | Alcohol consumption status of the customer |
| exercise | Regular exercise status of the customer |
| weight_change_in_last_one_year | How much variation has been seen in the weight of the customer in last year |
| fat_percentage | Fat percentage of the customer while applying the insurance |
| insurance_cost | Total Insurance cost |

# Modelling Approach

Step-1: Data preprocessing: Converted all ‘object’ datatype variables to numeric through one-hot encoding or categorizing them from 0-5.

Step-2: Data restructuring: Divided the 15 cities into four zones; North, East, South and West.

Step-3: Variance Inflation Factor was checked for all the features.

In this step, three variables namely ‘zone_West’, ‘occupation_Business’, ‘smoking_status_smokes’ were eliminated.

The final dataset was left with 27 features for modelling.

# Modelling Approach

- Reference for Zone assignment of the cities:

| location | zone | |
|---|---|---|
| Ahmedabad | North | 1677 |
| Bangalore | South | 1742 |
| Bhubaneswar | East | 1704 |
| Chennai | South | 1669 |
| Delhi | North | 1680 |
| Guwahati | East | 1672 |
| Jaipur | North | 1706 |
| Kanpur | North | 1664 |
| Kolkata | East | 1620 |
| Lucknow | North | 1637 |
| Mangalore | South | 1697 |
| Mumbai | West | 1658 |
| Nagpur | West | 1663 |
| Pune | West | 1622 |
| Surat | West | 1589 |

Name: count, dtype: int64

# Modelling Approach

- 1st stage: Choosing base models

The business wants to predict the *insurance cost* of an individual, which is a continuous variable. To build a predictive model where the target variable is continuous, Linear Regression can be used.

Along with Linear Regression, we also build three other models that manipulate the linear regression algorithm to generate optimized output. The three additional models were Ridge, Lasso, and Elastic Net. This is so that we can choose the best performing model for final use.

- 2nd stage: Using Ensemble Techniques

In this round, two models were used: Gradient Boosting Regressor and Bagging Regressor.

# Modelling Approach

- 3rd stage: Tuning the models

The models were re-built after changing a few key characteristics of the dataset.

a. The dataset train and test split was changed from 70:30 to 20:80.

b. The null values that were initially dropped were now treated with KNN imputer.

c. The outliers were left untreated.

d. Grid Search CV was used on the best performing base models for hyperparameter tuning.

# Model Selection

1. Two models have been shortlisted from the model comparison table: Elastic Net from the set of linear models and Gradient Boosting Regressor from the ensemble model set.
2. Elastic Net and Linear Regression models have performed similarly. However, Elastic Net was shortlisted as it uses the power of Ridge and Lasso Regression to regularize the base linear regression model.
3. Gradient Boosting Regressor has performed well while Bagging Regressor is overfitting. Its performance on the train dataset is high but drops significantly on the test dataset, especially for MAPE and RMSE metrics.
4. The upcoming table compares the R-squared (R2), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) of all the final models.

# Model Comparison Table

| Model Category | Model Name | R2 Train | R2 Test | RMSE Train | RMSE Test | MAPE Train | MAPE Test |
|---|---|---|---|---|---|---|---|
| Linear | Linear Regression | 0.945 | 0.944 | 3368 | 3374 | 0.15 | 0.16 |
| | Ridge | 0.945 | 0.944 | 3368 | 3374 | 0.389 | 0.394 |
| | Lasso | 0.945 | 0.944 | 3369 | 3373 | 0.39 | 0.394 |
| | Elastic Net | 0.945 | 0.944 | 3368 | 3374 | 0.152 | 0.155 |
| Ensemble | Gradient Boosting | 0.956 | 0.955 | 3013 | 3026 | 0.121 | 0.121 |
| | Bagging | 0.991 | 0.946 | 1355 | 3320 | 0.052 | 0.131 |

- Parameters for the best models:

**1. Gradient Boosting**

Loss = 'huber’ (default: 'squared error’)
n_estimators = 100 (default: 100)

**2. Elastic Net**

Alpha = 0.01 (default: 1.0)
L1 ratio = 0.3 (default: 0.5)

# Final Model Output I

- Elastic Net:

Insurance cost =

-79941.93 +

(-445.51 * regular_checkup_last_year) +

(181.69 * adventure_sports) +

(-55.83 * visited_doctor_last_1_year) +

(3.77 * age) +

(252.86 * heart_decs_history) +

(1489 * weight) +

(162.66 * weight_change_in_last_one_year) +

(1203 * covered_by_any_other_company)

# Business Insights and Recommendations I

- If an individual had a checkup last year, then her insurance cost will reduce by 446 units. Individuals that go for regular checkups are less likely to have high insurance costs, given that all the other variables remain constant.
- If an individual is involved in adventure sports, their insurance cost will go up by 182 units, given that all the other variables remain constant. Health risks in individuals who attempt adventure sports increase, leading to higher insurance costs.
- If an individual visited a doctor last year, it can bring down the insurance cost by 56 units, given that all the other variables remain constant. The more regular the checkups, the less chance of bigger health risks since big health problems could have an early detection.
- As an individual gets older, the insurance cost will increase 3.77 units, given that all the other variables remain constant.

# Business Insights and Recommendations I

- If an individual has history of a heart decease, the insurance cost is expected to increase by 253 units, given that all the other variables remain constant.
- A unit increase in weight will increase the insurance cost by 1489 units, given that all the other variables remain constant.
- If a person has experienced weight change in the last year, it could result in an increase of 163 units in the insurance cost, given that all the other variables remain constant.
- If a person is covered by any other company, the insurance cost can have an increase of 1203 units, given that all the other variables remain constant.

# Key Takeaways

1. According to the Elastic Net model, regular checkup last year is negatively correlated with the insurance cost. The lower the number of checkups, the higher the insurance cost for the individual.
2. The feature ‘visited_doctor_last_1_year’ is also negatively correlated to the insurance cost. The higher the number, the lower the insurance cost for the individual.
3. All the other variables are positively correlated with the target variable.

# Final Model Output II

- Gradient Boosting Regressor:

| Features | Imp |
| --- | --- |
| weight | 0.995487 |
| covered_by_any_other_company | 0.002108 |
| regular_checkup_last_year | 0.001511 |
| weight_change_in_last_one_year | 0.000511 |
| age | 0.000222 |
| visited_doctor_last_1_year | 0.000085 |
| adventure_sports | 0.000048 |
| heart_decs_history | 0.000028 |

# Business Insights and Recommendations II

- Through feature importance, the Gradient Boosting Regressor (GBR) showcases the influence of each feature over the target variable.
- As suggested by Elastic Net, the GBR model also used the ‘weight’ feature the most to predict the target variable, followed by ‘covered_by_any_other_company’ and ‘regular_checkup_last_year’.

# Final Suggestions for the Company

1. Under the ‘smoking_status’ category, the option *Unknown* can be removed (if it were a customer survey form that was used for data collection).
2. The company can add another feature named height along with ‘weight’. This will help us to calculate the ‘bmi’ field accurately, hence improving data integrity.
3. Individuals can be segmented based on the location to run specific campaigns depending on the feature specifications of the city.
4. The ‘year_last_admitted’ field can be made mandatory to avoid null values. This column had to dropped as it had around 48% null values. This feature has a strong correlation with the insurance cost (target) variable.
5. Make as many features mandatory to fill as possible to avoid null values.