HEALTH INSURANCE CLAIM PREDICTION

ADSP 32009 - Data Science in Healthcare (Autumn 23)

Mariam Adeyemo, Minh Vo

December 7, 2023

AGENDA

1. Problem Statement
2. Data Overview & Assumptions
3. Exploratory Data Analysis
4. Data Preprocessing
5. Modeling
6. Model Performance & Evaluation
7. Limitations & Future Work

PROBLEM STATEMENTS & OBJECTIVES

Insurance Claim

- Accurate premium setting is essential to match individual health risks.
- Inaccuracies in claim predictions can significantly impact financial outcomes.
- Complex health factors challenge accurate predictions.

Project Objectives

- Develop a data-driven solution for more precise health claim predictions.
- Integrate a range of health indicators to inform claim amount predictions.
- Support insurers' financial decision-making with data-driven accuracy.

Data Overview and Assumption

The insurance claim dataset provides a comprehensive examination of demographic and health data from insurance claims.

- Data source: Kaggle
- Initial total data: **15000 rows**
- Number of columns: **13 columns**
  - Policyholders' age
  - Gender
  - BMI (Body Mass Index)
  - Blood pressure levels
  - Smoking status
  - City
- Target variable: **Claim amount**

The final insurance claim dataset after data processing

- Final total data: **13904 rows**
- Number of columns: **25 columns**
  - **Age categories**
  - Gender
  - **BMI categories**
  - Diabetic status
  - Smoking status
  - **Region**

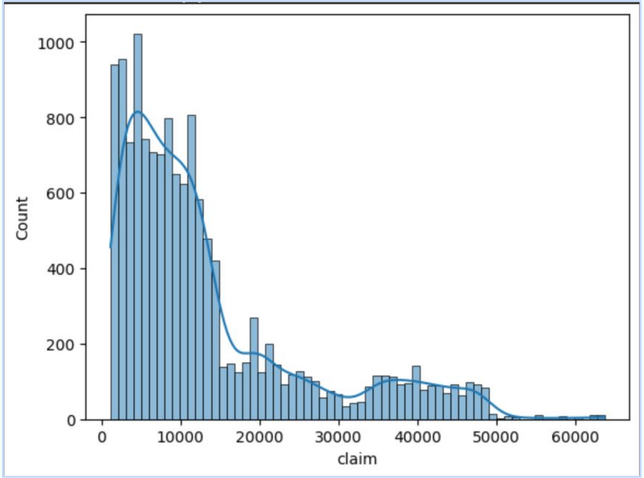
Target variable: **Claim amount**

Assumptions

- Diabetic and smoking status are likely associated with higher medical expenses
- Older individuals may incur higher insurance claims due to age-related health issues, while gender differences may reveal distinct health patterns

EXPLORATORY DATA ANALYSIS

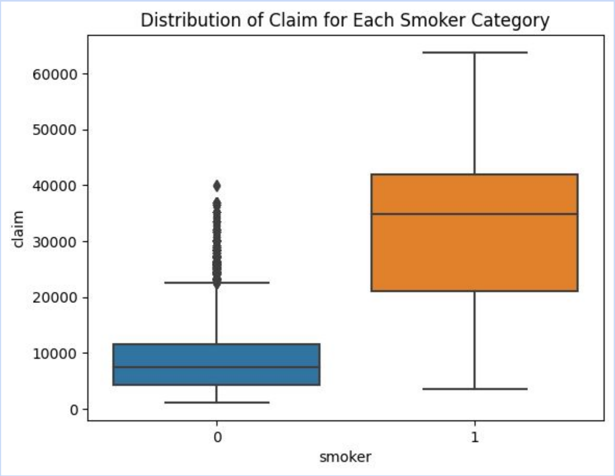
Target Variable



Claim Summary:

- Right skewed
- There are many outliers
- Claim amount is generally between 1,121 - 63,770
- Average claim payment is ~13,500

Smoker

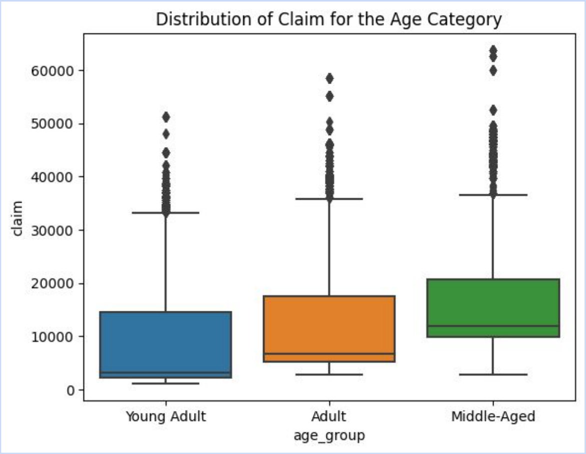


Smoker vs Claim Summary:

- Smokers have a high claim amount compared to non-smokers
- Average claim payment for smokers is 32,101
- Average claim payment for non-smokers is 8,745

EXPLORATORY DATA ANALYSIS

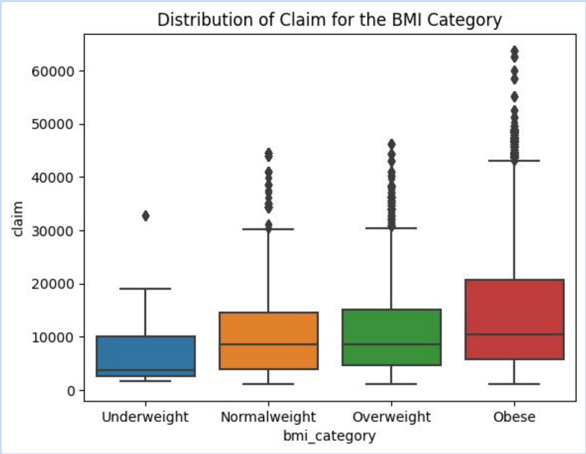
Age Group



Age Group vs Claim Summary:

- Middle-aged individuals typically have higher average claim amounts
- Average claim payment for middle-aged individuals is 17,099

BMI



BMI Category vs Claim Summary:

- Obese individuals show higher average claim amounts
- Average claim payment for obese individuals is 15,835

DATA PREPROCESSING

1

Data Cleaning

- Removed the duplicates values
- Imputed the missing values
- Checked for outliers

2

Feature Engineering

- Binned continuous features: age and BMI
- Converted the city column into regions
- Selected relevant features for modeling

3

Data Transformation

- Encoded all categorical variables
- Standardize features using StandardScaler()
- **Data Split:** Train set (80%), Test set (20%)

MODELING

01 **Linear Regression**

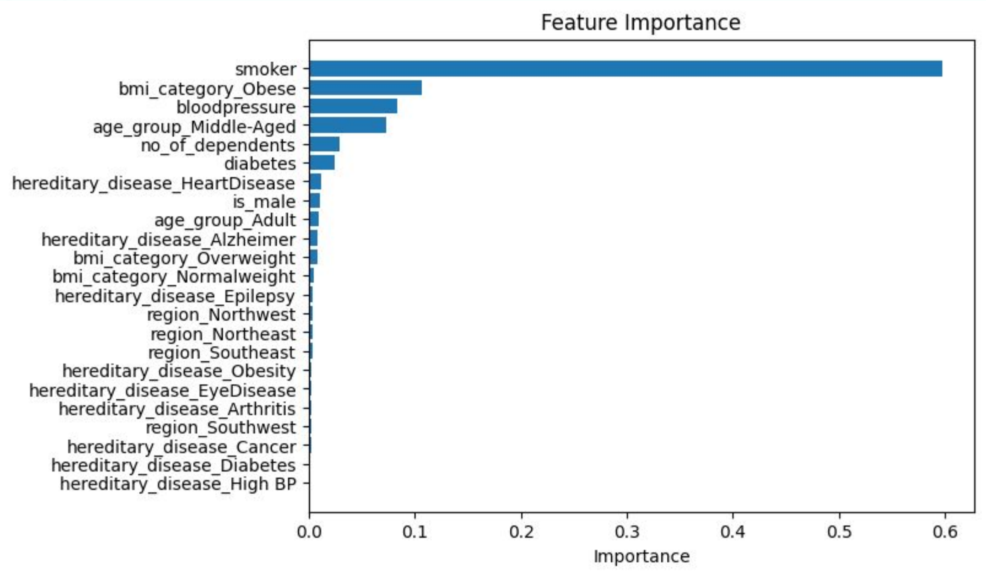
02 **Decision Tree**

03 **Random Forest**

04 **XGBoost**

- Linear Regression is considered based model.
- Conduct feature selection on Random Forest and XGBoost.
- Apply hyperparameter tuning on Decision Tree, Random Forest and XGBoost.

FEATURE SELECTION

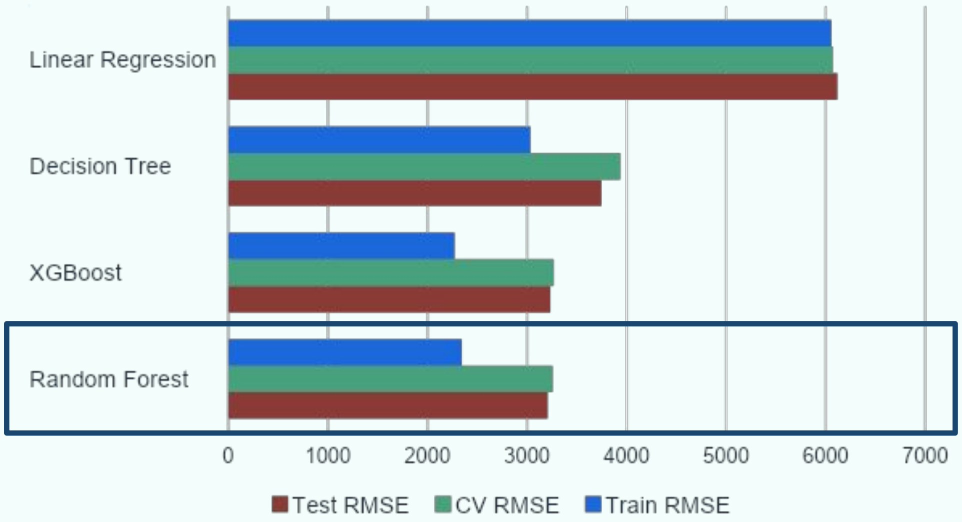


Selected features:

- smoker
- bmi_category_Obese
- blood_pressure
- age_group_Middle-Aged
- number_of_dependents
- diabetes
- hereditary_disease_HeartDisease
- is_male (gender)
- age_group_Adult
- hereditary_disease_Alzheimer
- bmi_category_Overweight

MODEL PERFORMANCE EVALUATION

Random Forest is identified as the best model for insurance claim amount prediction.



Model	Test RMSE	CV RMSE	Train RMSE
Random Forest	3207.04	3256.68	2344.85
XGBoost	3233.33	3267.1	2272.43
Decision Tree	3746.38	3935.82	3033.52
Linear Regression	6116.49	6069.14	6055.63

HEALTHCARE IMPACTS

Enhance Decision-Making

- ❏ Accurate Risk Assessment
- ❏ Financial Stability and Risk Management
- ❏ Efficient Resource Allocation

Stakeholder Benefits

- ❏ **Insurers:** reduce operational costs and minimize financial risks.
- ❏ **Policyholders:** enhance trust in insurance processes.
- ❏ **Healthcare Providers:** better planning for patient care needs.

LIMITATIONS & FUTURE WORK

Limitations

- ❑ The dataset may not capture the full spectrum of U.S. demographic and health.
- ❑ The presence of missing values and choice of imputation method can influence the model's accuracy.
- ❑ Health factors may evolve over time.

Future Work

- ❑ Implementing more advanced modeling techniques (e.g. Neural Networks).
- ❑ Exploring external data sources to enhance the model's performance and expands its scope.

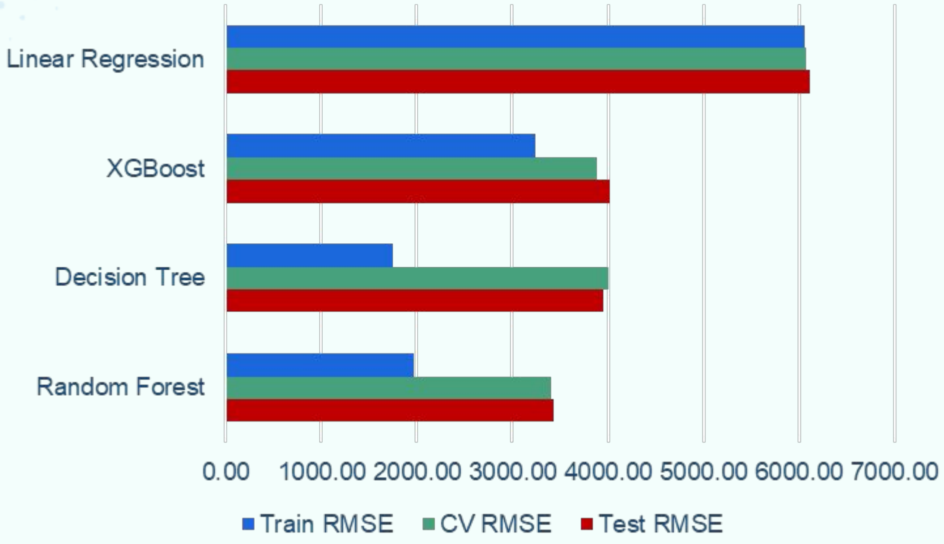
THANK YOU

APPENDIX

DATA DESCRIPTION

Feature	Description
age	Age of the policyholder
sex	Gender of policyholder
weight	Weight of the policyholder
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight
no_of_dependents	Number of dependent persons on the policyholder
smoker	Indicates policyholder is a smoker or a non-smoker (non-smoker=0;smoker=1)
claim	The amount claimed by the policyholder (Numeric)
bloodpressure	Blood pressure reading of policyholder (Numeric)
diabetes	Indicates policyholder suffers from diabetes or not (non-diabetic=0; diabetic=1)
regular_ex	A policyholder regularly exercises or not (no-exercise=0; exercise=1)
job_title	Job profile of the policyholder
city	The city in which the policyholder resides
hereditary_diseases	A policyholder is suffering from a hereditary disease or not

BASE MODEL PERFORMANCE



Model	Test RMSE	CV RMSE	Train RMSE
Random Forest	3433.78	3407.61	1969.98
Decision Tree	3957.48	4000.22	1748.39
XGBoost	4014.99	3887.43	3234.80
Linear Regression	6116.49	6069.14	6055.63

HYPERPARAMETER TUNING

Decision Tree:

DecisionTreeRegressor(max_depth=15, min_samples_leaf=5, min_samples_split=9)

Random Forest:

RandomForestRegressor(max_depth=18, max_features='sqrt', min_samples_split=3, n_estimators=400)

XGBoost:

XGBRegressor(base_score=None, booster=None, callbacks=None, colsample_bylevel=None, colsample_bynode=None, colsample_bytree=None, device=None, early_stopping_rounds=None, enable_categorical=False, eval_metric=None, feature_types=None, gamma=None, grow_policy=None, importance_type=None, interaction_constraints=None, learning_rate=0.01, max_bin=None, max_cat_threshold=None, max_cat_to_onehot=None, max_delta_step=None, max_depth=15, max_leaves=None, min_child_weight=None, n_estimators=400,...)