Health Insurance Premium Prediction

An Interactive ML App with Real-Time Predictions

Presented by: Neethu Manikantan

PROBLEM STATEMENT

- Rising healthcare costs make fair premium estimation essential
- Insurance companies need data-driven pricing models
- Users expect transparency in premium determination

PROJECT OBJECTIVE

- Help insurance companies estimate premiums efficiently.
- Develop a predictive model to predict health insurance premium using ML
- Use demographic and medical features
- Provide real-time predictions via Streamlit app

BUSINESS REQUIREMENTS

- Develop a high-accuracy (>97%) predictive model to predict health insurance premium using ML
- The percentage difference between the predicted and actual value on a minimum of 95% of the errors should be less than 10%
- Deploy the model in the cloud so that an insurance companies can run it from anywhere
- Create an interactive Streamlit application that insurance companies can use for predictions

DATA COLLECTION

Dataset (~50000 records)

Feature Name	Description
age	Age of the individual
gender	Gender: Male / Female
region	Geographic location: Northwest / Southeast / Northeast / Southwest
marital_status	Marital status: Unmarried / Married
number_of_dependants	Count of dependents
bmi_category	BMI category: Underweight / Normal / Overweight / Obesity
smoking_status	Smoking habit: No Smoking / Regular / Occasional
employment_status	Employment type: Salaried / Freelancer / Self-Employed
income_level	Income group: <10L / 10L–25L / 25L–40L / >40L
income_lakhs	Income in lakhs (numerical value)
medical_history	Details of past medical conditions -'Diabetes' 'High blood pressure' 'No Disease' 'Diabetes & High blood pressure' 'Thyroid' 'Heart disease' 'High blood pressure & Heart disease' 'Diabetes & Thyroid' 'Diabetes & Heart disease'
insurance_plan	Type of plan: Bronze / Silver / Gold
annual_premium_amount	**Target variable**: Premium amount to be predicted

EXPLORATORY DATA ANALYSIS

- Missing value handling
    - remove null values
    - remove duplicate rows
- Handling Invalid Data
    - replace negative number of dependents with absolute value
- Numerical Column Analysis
    - Univariate Analysis: box plot
        - Age: limit set to 100 removed greater values
        - income : used 99.9th percentile as upper bound as per business requirements
    - Bivariate Analysis:
        - No major insights
- Categorical Columns Analysis:
    - Univariate:
        - clean smoking_status values to unique values
    - Bivariate:
        - bar plots: no major insights

FEATURE ENGINEERING

- Assign numerical values to medical history to form new column - normalized_risk_score
  - medical_history->disease1+disease2->assign scores->normalise scores
- Label encoding of ordinal features
  - insurance_plan = 'Bronze': 1, 'Silver': 2, 'Gold': 3
  - 'income_level = <10L':1, '10L - 25L': 2, '25L - 40L':3, '> 40L':4
- One hot encoding of nominal features
- Drop original columns from which new columns were derived= medical_history','disease1', 'disease2', 'total_risk_score

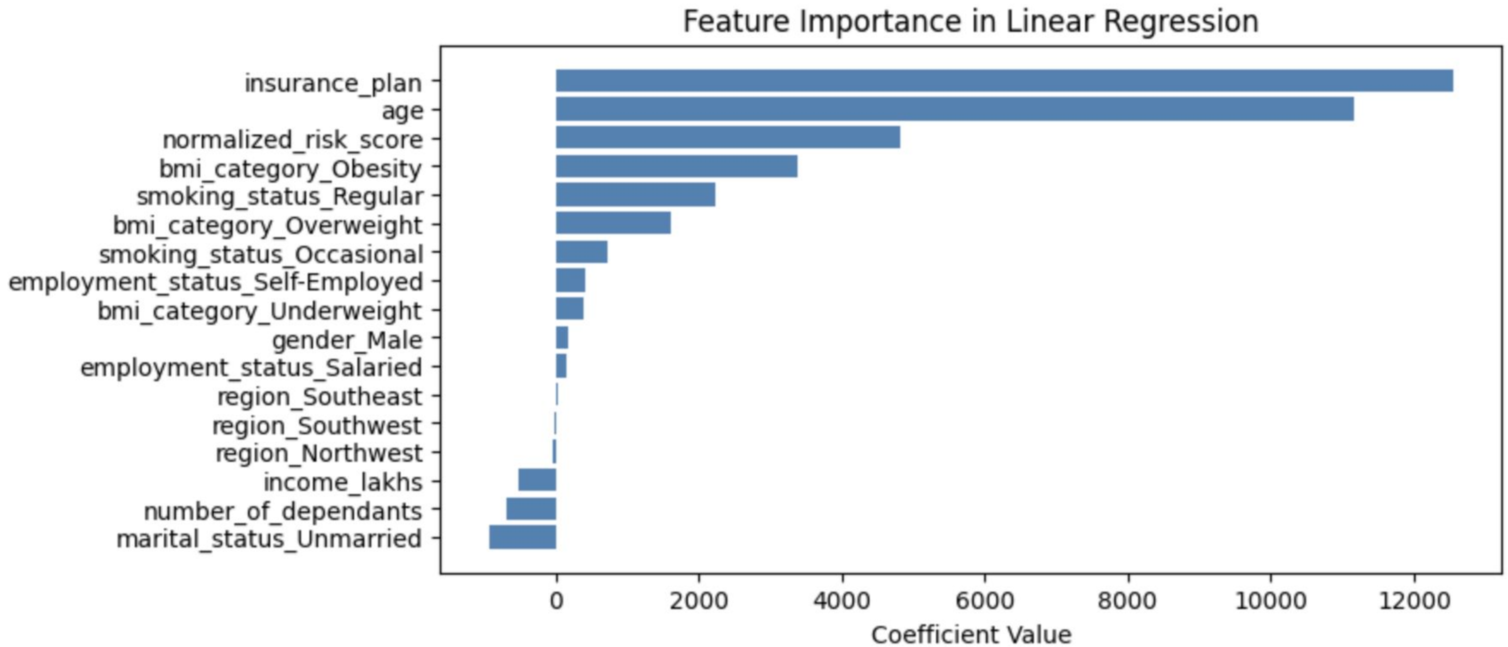
- Scaling the features using Min-Max Scaler
- Check Multicolinearity using VIF(Variance Inflation Factor)
    - Drop columns with VIF> 10
        - income_level

	Column	VIF
0	age	4.545825
1	number_of_dependants	4.526598
2	income_lakhs	2.480563
3	insurance_plan	3.445682
4	normalized_risk_score	2.687326
5	gender_Male	2.409980
6	region_Northwest	2.100789
7	region_Southeast	2.919775
8	region_Southwest	2.668314
9	marital_status_Unmarried	3.393718
10	bmi_category_Obesity	1.352748
11	bmi_category_Overweight	1.549907
12	bmi_category_Underweight	1.302636
13	smoking_status_Occasional	1.272744
14	smoking_status_Regular	1.777024
15	employment_status_Salaried	2.374628
16	employment_status_Self-Employed	2.132810

MODEL TRAINING

Linear regression

- MSE: 5165611.913027982
- RMSE: 2272.798256121291
- R2-score: 0.92805

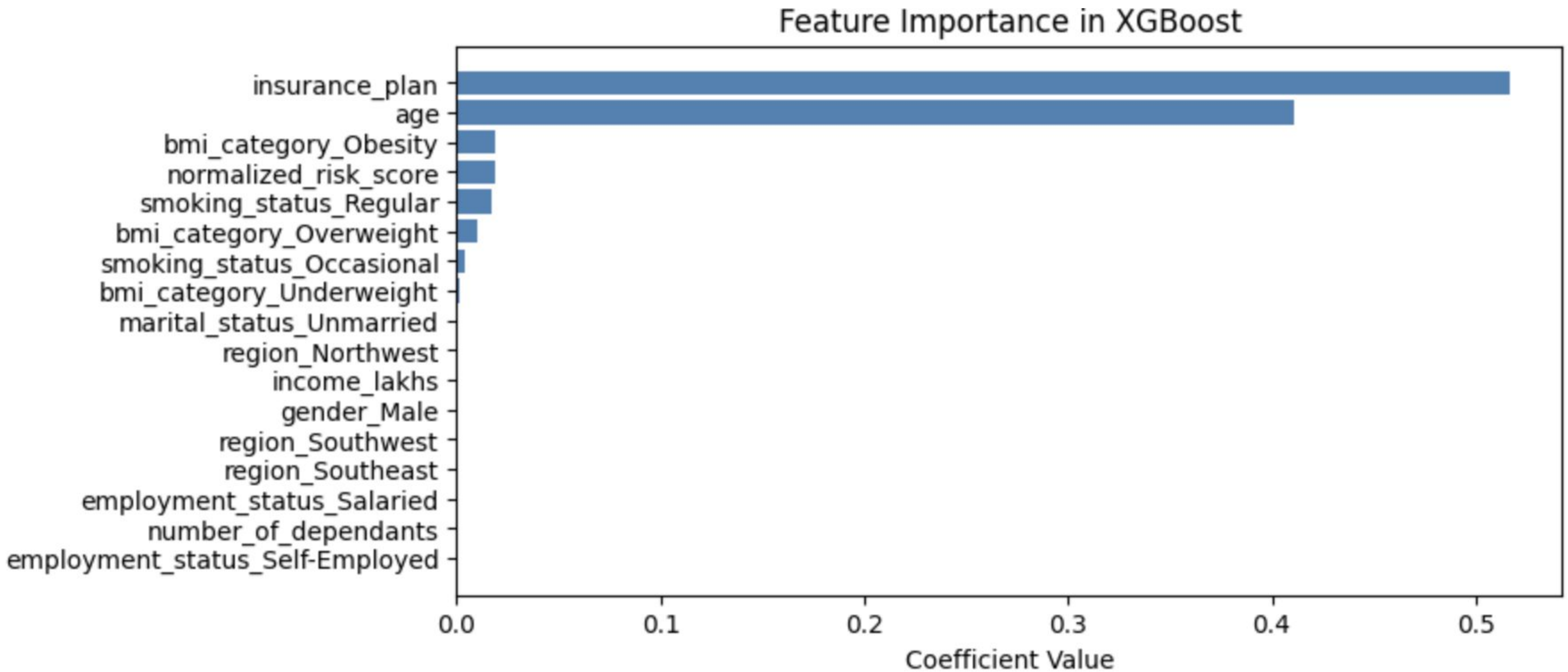


Ridge Regression Model

- MSE: 5165652.017016523
- RMSE: 2272.8070787060924
- R2-score: 0.928

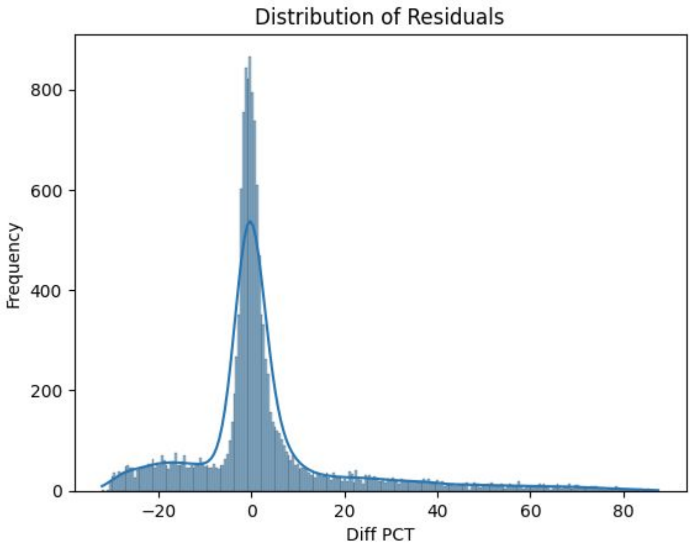
XGBoost

- MSE: 1563064.1356043513
- RMSE: 1250.2256338774819
- R2-score: 0.978



Checking business requirement

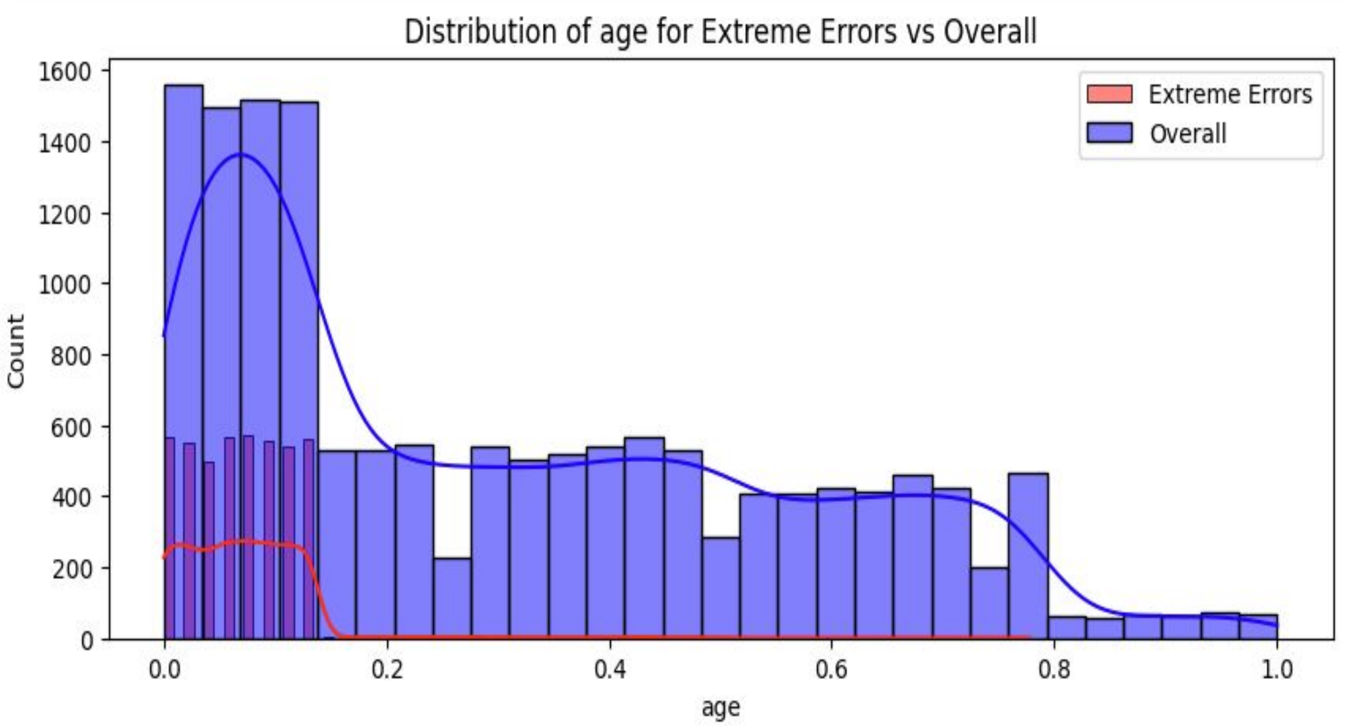
- Calculate residual percentage = (residual/y_test)*100
- residual = y_pred-y_test
- Set extreme_error_threshold = 10
- For 30% customers the model will either overcharge or undercharge by 10% or more



	actual	predicted	diff	diff_pct
42730	5018	7352.829590	2334.829590	46.529087
20029	5140	6670.849121	1530.849121	29.783057
4294	9631	7053.477539	-2577.522461	-26.762771
44419	4687	6670.849121	1983.849121	42.326629
6707	8826	10047.326172	1221.326172	13.837822

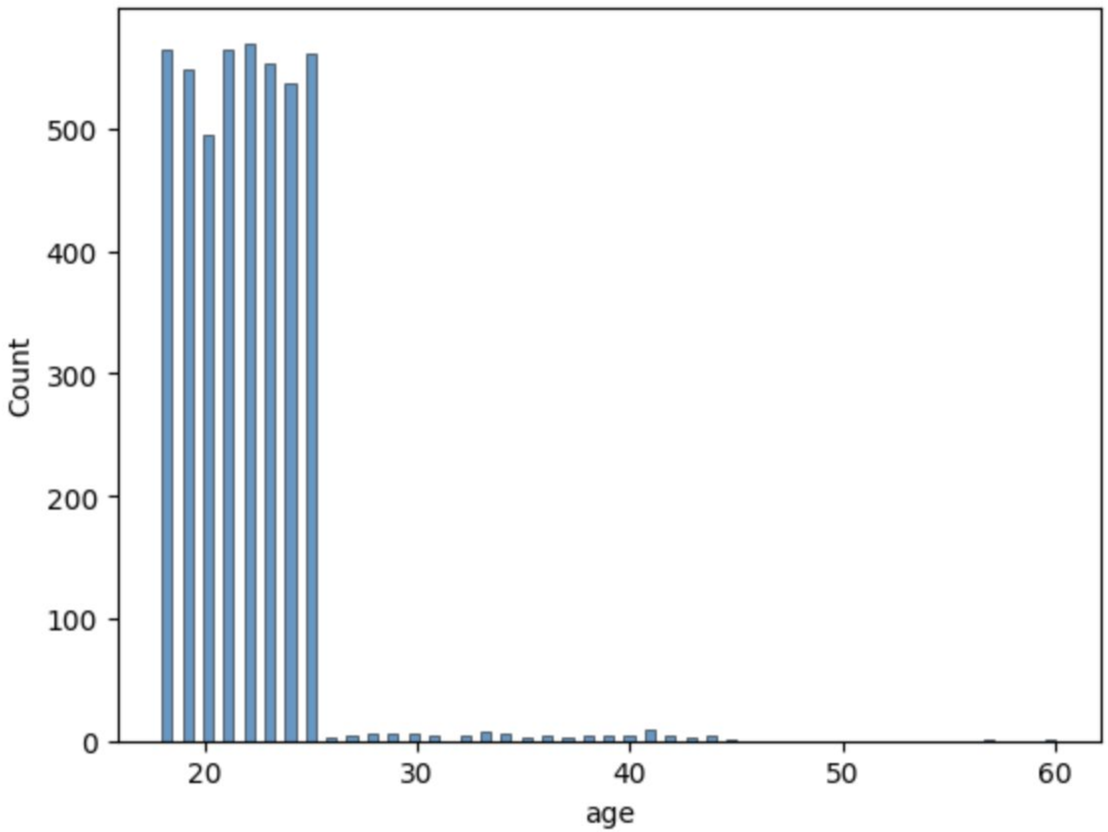
kde plot of all features with extreme errors

- found a pattern in age vs extreme errors
- majority of the extreme errors are coming from young age group



Age distribution in extreme errors list

- This shows errors are extreme for records with <25 years of age.
- We need to may be build a separate model for this segment



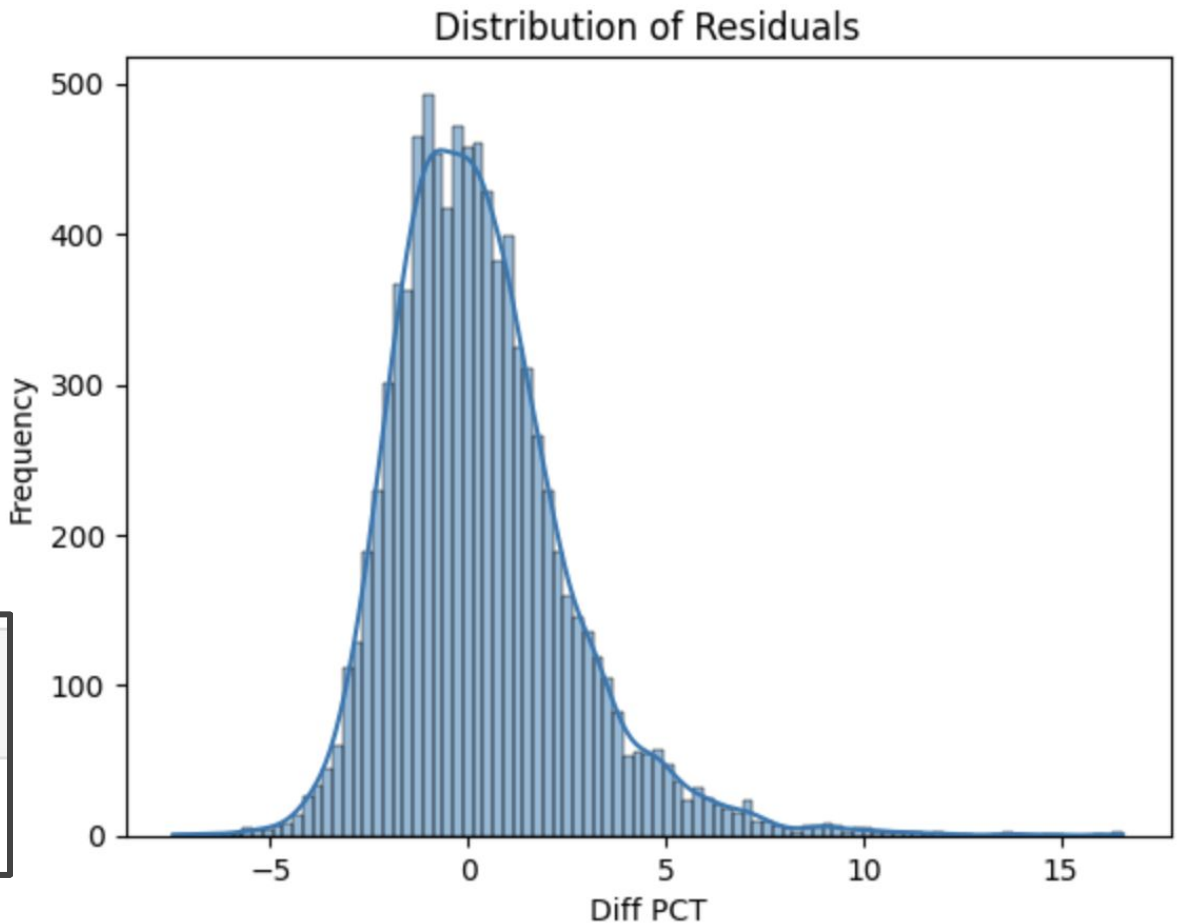
MODEL SEGMENTATION

Segment 1: Age>25

We have very few extreme errors (only 0.3%) which means this model looks good and no further investigation is required

```
extreme_results_df.shape
```

```
(29, 4)
```

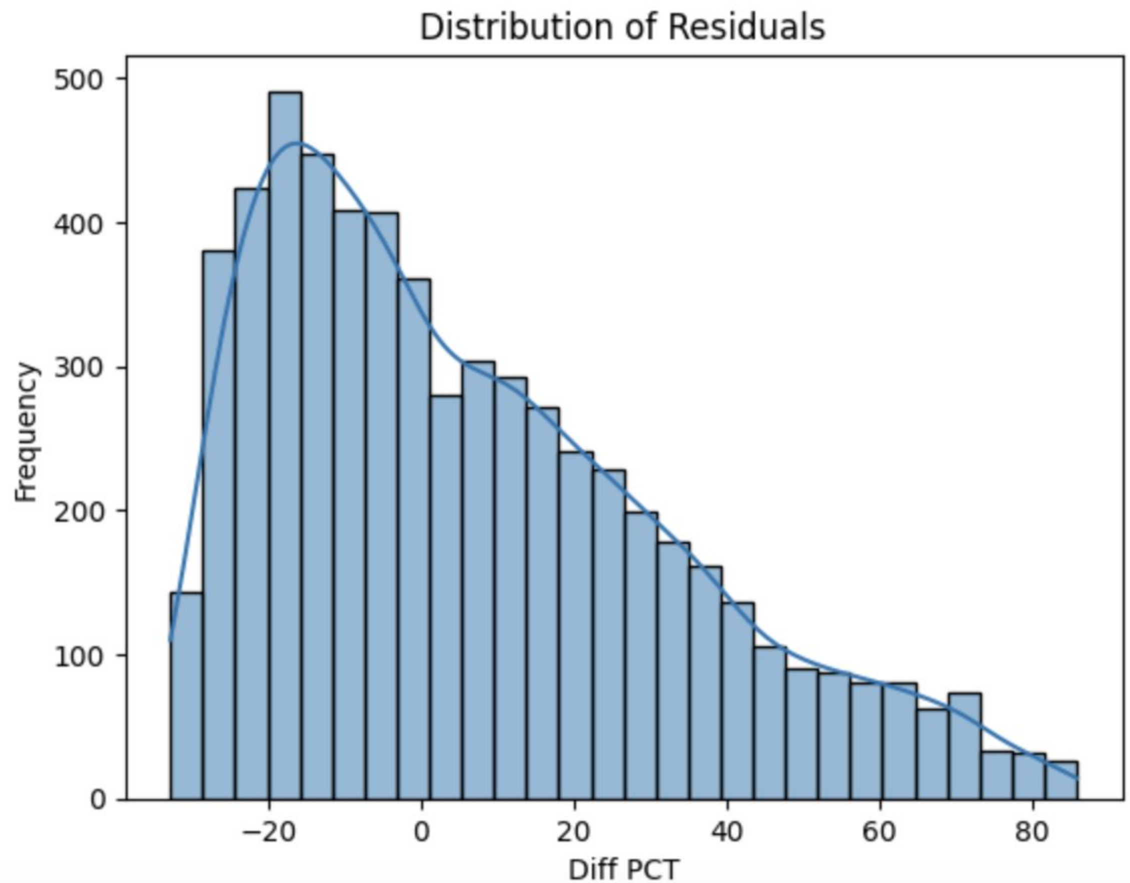


Segment 2: Age<25

- In this segment, we have 73% extreme errors.
- By comparing distributions of extreme errors vs features, we don't get much insights.
- May be we need more features in order to improve the performance

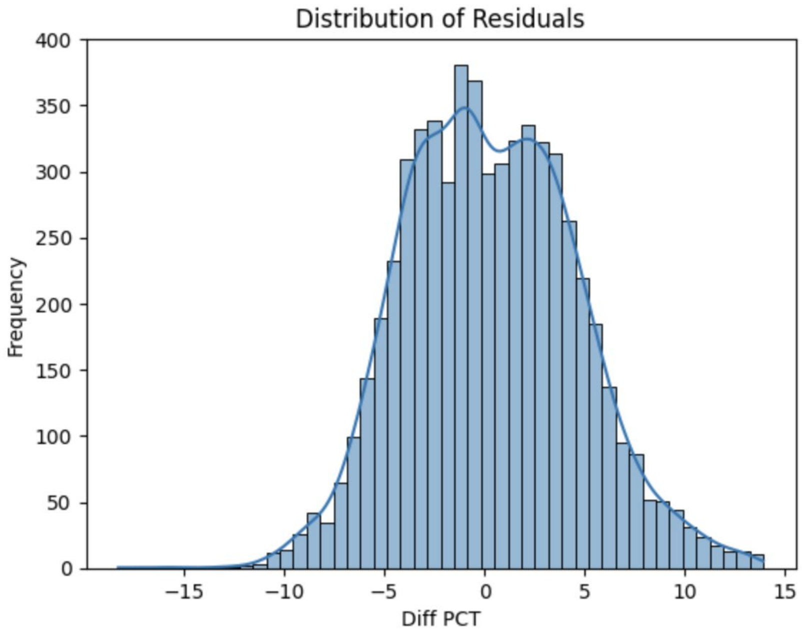
```
extreme_results_df.shape
```

```
(4404, 4)
```



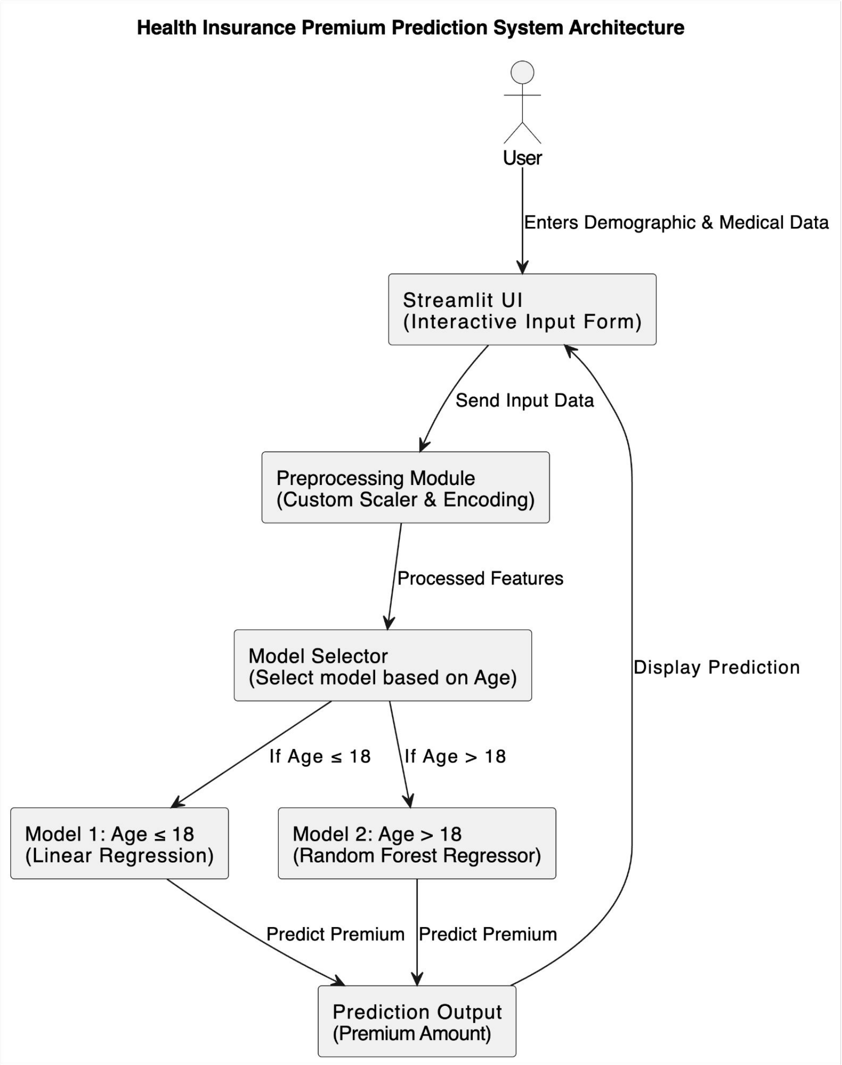
Adding new feature - Genetic Risk

- Added genetic risk feature
- Retrained both models
- Evaluation metric: R2-score:
  - Linear regression - 0.988
  - Ridge regression - 0.988
  - xgboost - 0.987



- Final Model
  - Linear regression-model explainability
- Extreme errors - 2%

System Architecture



FRONT END

Interactive Streamlit Application

- Real-time input via web interface
- Age-based prediction flow
- User-friendly frontend with form inputs and result display.

Health Insurance Prediction App

Age

19

Number of Dependants

0

Income in Lakhs

200

Genetical Risk

0

Insurance Plan

Silver

Employment Status

Salaried

Gender

Male

Marital Status

Unmarried

BMI Category

Obesity

Smoking Status

No Smoking

Region

Southeast

Medical History

No Disease

Predict

Predicted Health Insurance Cost: 8309

THANKYOU!