Baseline Notebook

Setup Environment

In [1]:

```
# DO NOT MODIFY THE CODE IN THIS CELL
!pip install -q utstd

from utstd.folders import *
from utstd.ipyrenders import *

at = AtFolder(
    course_code=36106,
    assignment="AT1",
)
at.run()

import warnings
warnings.simplefilter(action='ignore')
```

```
ERROR: Could not install packages due to an OSError: [WinError 5] 拒绝访问。: 'C:\\Users\\broha
o\\AppData\\Local\\Programs\\Python\\Python311\\Lib\\site-packages\\~0learn\\.libs\\msvcp140.d
ll'
Consider using the `--user` option or check the permissions.


[notice] A new release of pip available: 22.3.1 -> 25.2
[notice] To update, run: python.exe -m pip install --upgrade pip
You can now save your data files in: c:\Users\brohao\Desktop\UTS\36106\AT1\36106\assignment\AT
1\data
```

Student Information

In [2]:

```
student_name = "Jiayu Hao"
student_id = "25948860"
```

In [3]:

```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h1", key='student_name', value=student_name)
```

student_name

Jiayu Hao

In [4]:

```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h1", key='student_id', value=student_id)
```

student_id

25948860

---

0. Python Packages

0.a Install Additional Packages

> If you are using additional packages, you need to install them here using the command: `! pip install <package_name>`

In [5]:

```
!pip install scikit-learn
!pip install numpy
```

```
Requirement already satisfied: scikit-learn in c:\users\brohao\appdata\local\programs\python\p
ython311\lib\site-packages (1.6.1)
Requirement already satisfied: numpy>=1.19.5 in c:\users\brohao\appdata\local\programs\python
\python311\lib\site-packages (from scikit-learn) (2.3.2)
Requirement already satisfied: scipy>=1.6.0 in c:\users\brohao\appdata\local\programs\python\p
ython311\lib\site-packages (from scikit-learn) (1.16.1)
Requirement already satisfied: joblib>=1.2.0 in c:\users\brohao\appdata\local\programs\python
\python311\lib\site-packages (from scikit-learn) (1.5.1)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\brohao\appdata\local\programs
\python\python311\lib\site-packages (from scikit-learn) (3.6.0)
```

```
[notice] A new release of pip available: 22.3.1 -> 25.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
Requirement already satisfied: numpy in c:\users\brohao\appdata\local\programs\python\python31
1\lib\site-packages (2.3.2)
```

```
[notice] A new release of pip available: 22.3.1 -> 25.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

0.b Import Packages

In [6]:

```
import pandas as pd
import altair as alt
```

In [7]:

```
import numpy as np
from sklearn.dummy import DummyRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

A. Assess Baseline Model

In [8]:

```
# DO NOT MODIFY THE CODE IN THIS CELL
# Load data
try:
  X_train = pd.read_csv(at.folder_path / 'X_train.csv')
  y_train = pd.read_csv(at.folder_path / 'y_train.csv')

  X_val = pd.read_csv(at.folder_path / 'X_val.csv')
  y_val = pd.read_csv(at.folder_path / 'y_val.csv')

  X_test = pd.read_csv(at.folder_path / 'X_test.csv')
  y_test = pd.read_csv(at.folder_path / 'y_test.csv')
except Exception as e:
  print(e)
```

A.1 Generate Predictions with Baseline Model

In [9]:

```
# Predict mean value of train set
baseline = DummyRegressor(strategy="mean")
```

```
baseline.fit(X_train, y_train)

y_train_pred = baseline.predict(X_train)
y_val_pred   = baseline.predict(X_val)
```

A.2 Selection of Performance Metrics

> Provide some explanations on why you believe the performance metrics you chose is appropriate

In [10]:

```
# MAE is the main metric
def evaluate(y_true, y_pred, dataset=""):
    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    r2 = r2_score(y_true, y_pred)
    print(f"--- {dataset} ---")
    print(f"MAE : {mae:.2f}")
    print(f"RMSE: {rmse:.2f}")
    print(f"R²  : {r2:.3f}")
```

In [11]:

```
# Provide some explanations on why you believe the performance metrics you chose is appropria
performance_metrics_explanations = """
It is important to use MAE as the main metric because it directly shows the average difference
MAE is less sensitive to extreme values, which makes it more suitable for insurance pricing.
RMSE and R² are also reported to provide additional insight.
"""
```

In [12]:

```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='performance_metrics_explanations', value=performance_metrics_expla
```

performance_metrics_explanations

It is important to use MAE as the main metric because it directly shows the average difference between predicted and actual premiums. MAE is less sensitive to extreme values, which makes it more suitable for insurance pricing. RMSE and R² are also reported to provide additional insight.

A.3 Baseline Model Performance

> Provide some explanations on model performance

In [13]:

```
# Baseline Model Performance(A.2 defined)
print("Baseline Model Performance:")
evaluate(y_train, y_train_pred, dataset="Train")
evaluate(y_val, y_val_pred, dataset="Validation")
```

```
Baseline Model Performance:
--- Train ---
MAE : 38.09
RMSE: 44.99
R²  : 0.000
--- Validation ---
MAE : 147.72
RMSE: 160.95
R²  : -0.456
```

In [14]:

```
# Provide some explanations on model performance
baseline_performance_explanations = """
The baseline model shows very low error on the training set but much higher error on the vali
The result confirms that predicting only with simple averages is not reliable and better mode
"""
```

In [15]:

```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='baseline_performance_explanations', value=baseline_performance_exp
```

baseline_performance_explanations

The baseline model shows very low error on the training set but much higher error on the validation set. The result confirms that predicting only with simple averages is not reliable and better models are needed for premium pricing.