Experiment 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\\brohao\\AppData\\Local\\Programs\\Python\\Python311\\Lib\\site-packages\\~1learn\\.libs\\msvcp140.dll'
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\AT1\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 numpy
!pip install scikit-learn
```

```
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
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
```

0.b Import Packages

In [6]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
import pandas as pd
import altair as alt
```

In [7]:
```
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_absolute_error
import numpy as np
```

A. Experiment Description

In [8]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
experiment_id = "1"
print_tile(size="h1", key='experiment_id', value=experiment_id)
```

experiment_id

1

In [9]:
```
# Present the hypothesis you want to test, the question you want to answer or the insight you
# Explain the reasons why you think it is worthwhile considering it
experiment_hypothesis = """
The hypothesis is that regularized linear regression can reduce overfitting risk and improve 
The question is how different alpha values affect model performance.
This is worthwhile because if Ridge regression shows clear improvement over the baseline (MAE
"""
```

In [10]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='experiment_hypothesis', value=experiment_hypothesis)
```

experiment_hypothesis

The hypothesis is that regularized linear regression can reduce overfitting risk and improve generalization in premium prediction. The question is how different alpha values affect model performance. This is worthwhile because if Ridge regression shows clear improvement over the baseline (MAE:148), it proves that simple linear methods can already capture useful patterns and provide value for pricing decisions.

In [11]:
```
# Detail what will be the expected outcome of the experiment. If possible, estimate the goal y
# List the possible scenarios resulting from this experiment.
experiment_expectations = """
The expected outcome is that alpha will influence the stability of MAE in Ridge regression, sh
If validation MAE drops clearly compared to the baseline of 148, it means the model are usefu
If performance remains weak, it suggests the need for more complex models or better feature se
"""
```

In [12]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='experiment_expectations', value=experiment_expectations)
```

experiment_expectations

The expected outcome is that alpha will influence the stability of MAE in Ridge regression, showing how regularization strength affects performance. If validation MAE drops clearly compared to the baseline of 148, it means the model are useful for prediction. If performance remains weak, it suggests the need for more complex models or better feature selection.

B. Feature Selection

In [13]:
```
# 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)
```

In [14]:
```
features_list = list(X_train.columns)
print("Number of features:", len(features_list))
```

```
Number of features: 34
```

In [15]:
```
# Provide a rationale on why you are selected these features but also why you decided to remo
feature_selection_explanations = """
The selected features include numerical variables such as customer seniority, vehicle attribu
```

```
Identifiers such as ID, name, address, phone, and email were removed because they do not contri
Only features with direct or indirect impact on premiums were kept.
"""
```

In [16]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='feature_selection_explanations', value=feature_selection_explanati
```

feature_selection_explanations

The selected features include numerical variables such as customer seniority, vehicle attributes, and claim history, as well as one-hot encoded categorical variables like gender, policy type, and channel. Identifiers such as ID, name, address, phone, and email were removed because they do not contribute to prediction. Only features with direct or indirect impact on premiums were kept.

C. Train Machine Learning Model

C.1 Import Algorithm

In [17]:
```
# Provide some explanations on why you believe this algorithm is a good fit
algorithm_selection_explanations = """
Ridge regression is more stable than ordinary linear regression and is a good first choice fo
It can handle multicollinearity, such as correlations among claim-related features.
"""
```

In [18]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='algorithm_selection_explanations', value=algorithm_selection_expla
```

algorithm_selection_explanations

Ridge regression is more stable than ordinary linear regression and is a good first choice for testing regularization. It can handle multicollinearity, such as correlations among claim-related features.

C.2 Set Hyperparameters

In [19]:
```
# Set Hyperparameters
alphas = [0.01, 0.1, 1, 10, 100, 1000]
```

In [20]:
```
# Explain why you are tuning these hyperparameters
hyperparameters_selection_explanations = """
Alpha controls the strength of regularization.
A small alpha makes the model close to linear regression and may overfit, while a large alpha
The goal is to find a balanced alpha that gives stable and reliable results.
"""
```

In [21]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='hyperparameters_selection_explanations', value=hyperparameters_sele
```

Alpha controls the strength of regularization. A small alpha makes the model close to linear regression and may overfit, while a large alpha shrinks coefficients more and may underfit. The goal is to find a balanced alpha that gives stable and reliable results.

C.3 Fit Model

In [22]:
```
results = []

# for each alpha
for alpha in alphas:
    model = Ridge(alpha=alpha, random_state=42)
    model.fit(X_train, y_train)

    # training set prediction
    y_train_pred = model.predict(X_train)
    train_mae = mean_absolute_error(y_train, y_train_pred)

    # validation set prediction
    y_val_pred = model.predict(X_val)
    val_mae = mean_absolute_error(y_val, y_val_pred)

    results.append((alpha, train_mae, val_mae))

# output results
print("Ridge Regression Results:")
print("alpha | Train MAE | Validation MAE")
for r in results:
    print(f"{r[0]:5} | {r[1]:9.2f} | {r[2]:14.2f}")
```

```
Ridge Regression Results:
alpha | Train MAE | Validation MAE
 0.01 |     34.67 |         127.52
  0.1 |     34.67 |         127.52
    1 |     34.67 |         127.52
   10 |     34.68 |         127.52
  100 |     34.70 |         127.33
 1000 |     34.78 |         126.95
```

D. Model Evaluation

D.1 Model Technical Performance

In [23]:
```
# Provide some explanations on model performance
model_performance_explanations = """
The Ridge regression shows stable MAE across all alpha values, with validation MAE around 127
This is better than the baseline of 148 but only a small improvement.
The results suggest that linear relationships exist, but regularization strength has little e
"""
```

In [24]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='model_performance_explanations', value=model_performance_explanati
```

model_performance_explanations

The Ridge regression shows stable MAE across all alpha values, with validation MAE around 127. This is better than the baseline of 148 but only a small improvement. The results suggest that linear relationships exist, but regularization strength has little effect, and more advanced models may be needed for further gains.

D.2 Business Impact from Current Model Performance

In [25]:
```
# Interpret the results of the experiments related to the business objective set earlier. Est
business_impacts_explanations = """
A lower MAE means premium predictions are closer to real values, helping the company set faire
The current MAE is still high, so errors remain significant.
Overestimation may lead to customer loss, while underestimation may expose the company to higl
"""
```

In [26]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h3", key='business_impacts_explanations', value=business_impacts_explanation
```

business_impacts_explanations

A lower MAE means premium predictions are closer to real values, helping the company set fairer prices. The current MAE is still high, so errors remain significant. Overestimation may lead to customer loss, while underestimation may expose the company to higher claim costs.

E. Conclusion

In [27]:
```
experiment_outcome = "Hypothesis Partially Confirmed"
```

In [28]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h2", key='experiment_outcomes_explanations', value=experiment_outcome)
```

experiment_outcomes_explanations

Hypothesis Partially Confirmed

In [29]:
```
# Reflect on the outcome of the experiment and list the new insights you gained from it. Prov
# Given the results achieved and the overall objective of the project, list the potential nex

experiment_results_explanations = """
Ridge regression provided a useful benchmark but did not achieve the level of accuracy needed
The improvement is limited, and regularization strength (alpha) had little effect, suggesting
Next steps and expected uplift:
Lasso Regression may perform feature selection and remove weak variables, improving interpret
KNN Regression can test nonlinearity and see if complex patterns between customer and vehicle
Feature Engineering is to add interaction terms (e.g., car_age * vehicle_value) or other risk
"""
```

In [30]:
```
# DO NOT MODIFY THE CODE IN THIS CELL
print_tile(size="h2", key='experiment_results_explanations', value=experiment_results_explana
```

Ridge regression provided a useful benchmark but did not achieve the level of accuracy needed for production. The improvement is limited, and regularization strength (alpha) had little effect, suggesting that Ridge regression cannot fully capture the complexity of the problem. Next steps and expected uplift: Lasso Regression may perform feature selection and remove weak variables, improving interpretability and possibly reducing noise. KNN Regression can test nonlinearity and see if complex patterns between customer and vehicle features and premiums can be captured. Feature Engineering is to add interaction terms (e.g., car_age * vehicle_value) or other risk indicators that reflect real pricing factors.