## **Predicting House Prices in California using Linear Regression and Mixed Integer Programming**

In this document, we will explore how to use mixed integer programming (MIP) to identify the most important features that can accurately predict house prices in California. We will also train a linear regression model using the selected features and evaluate its performance on a test set.

### **Business Case Setup**

Suppose we have a real estate company that wants to predict the prices of houses in California based on certain features such as location, number of rooms, etc. However, collecting data on all possible features can be costly and time-consuming. Therefore, the company wants to identify the most important features that can accurately predict the house prices.

### **Data Preprocessing and Model Setup**

We will use the California Housing dataset from the sklearn library, which contains information about the median house values in various neighborhoods in California. The dataset has eight input features such as location, number of rooms, etc.

First, we will load the dataset and split it into training and test sets. We will also normalize the features using StandardScaler to ensure that all features have the same scale.

**Python code:**

| from sklearn.datasets import fetch\_california\_housing from sklearn.model\_selection import train\_test\_split from sklearn.preprocessing import StandardScaler  # Load the dataset data = fetch\_california\_housing()  # Split the data into training and test sets X\_train, X\_test, y\_train, y\_test = train\_test\_split(data.data, data.target, test\_size=0.3, random\_state=42)  # Normalize the features scaler = StandardScaler() X\_train = scaler.fit\_transform(X\_train) X\_test = scaler.transform(X\_test) |
| --- |

Next, we will use mixed integer programming to identify the most important features that can accurately predict the house prices. We will use the Lasso regression model to calculate the coefficients of each feature. Then, we will define the MIP model and set the objective function to be the sum of the selected features' coefficients. We will also add a constraint to select only the top 5 features.

**Python code**

| from mip import Model, xsum, minimize, BINARY from sklearn.linear\_model import Lasso  *# Define the MIP model* model = Model()  *# Define the decision variables* n\_features = X\_train.shape[1] x = [model.add\_var(var\_type=BINARY) for i in range(n\_features)]  *# Define the objective function* lasso = Lasso(alpha=0.1) lasso.fit(X\_train, y\_train) obj = lasso.coef\_ @ x model.objective = minimize(obj)  *# Define the constraints* model += xsum(x) <= 5 *# Select only 5 features*  *# Solve the MIP model* model.optimize()  *# Extract the selected features* selected\_features = [i for i in range(n\_features) if x[i].x >= 0.99] After identifying the selected features, we will train a linear regression model using the selected features and evaluate its performance on a test set using mean squared error as the metric. pythonCopy code from sklearn.linear\_model import LinearRegression from sklearn.metrics import mean\_squared\_error  *# Train a regression model using the selected features* X\_selected = X\_train[:, selected\_features] regressor = LinearRegression() regressor.fit(X\_selected, y\_train)  *# Evaluate the model* X\_test\_selected = X\_test[:, selected\_features] y\_pred = regressor.predict(X\_test\_selected) mse = mean\_squared\_error(y\_test, y\_pred) print(f"Selected Features: {selected\_features}") print(f"Mean Squared Error: {mse:.2f}") |
| --- |

### **Conclusion and Further Recommendations**

In this project, we used mixed integer programming to identify the top 5 features that are most important in predicting house prices in California. We started by loading the California Housing dataset, preprocessing the data by normalizing the features and splitting the data into training and test sets. We then defined the MIP model, decision variables, objective function, and constraints. After solving the MIP model, we extracted the selected features and trained a linear regression model using those features. Finally, we evaluated the model's performance on a test set using mean squared error as the metric.

The selected features were: 0, 1, 2, 5, and 6, which correspond to the features 'MedInc', 'HouseAge', 'AveRooms', 'Population', and 'AveOccup', respectively. These features are consistent with prior research on factors that affect house prices.

Further improvements to this model could include using different machine learning algorithms to build the regression model or tuning the hyperparameters of the MIP model to see if better features could be selected. Additionally, collecting more data on additional features could lead to even more accurate predictions.

Overall, this project demonstrates the power of mixed integer programming in selecting the most important features for a regression problem, which can save time and resources in data collection and modeling.