# Create a linear regression AI prediction using BQ ML

Implementing churn prediction using BigQuery ML involves several steps. Here's a step-by-step guide to help you through the process:

### Step 1: Prepare Your Data

1. **Load your data**: Ensure that your data is loaded into a BigQuery table. The data should include customer behavior features and a churn indicator (1 for churned, 0 for not churned).
2. **Clean your data**: Make sure the data is clean, with no missing or inconsistent values.

### Step 2: Feature Engineering

1. **Create features**: Generate features that might influence customer churn, such as the number of purchases, total spend, recency of last purchase, and every behaviour act your can think of such as ( call to customer success, emails to company ) etc.
2. **Split data**: Split your dataset into training and evaluation sets. For example, you might use 80% of your data for training and 20% for evaluation.

### Step 3: Create and Train the Model

1. **Create a training model**: Use SQL to create and train a logistic regression model in BigQuery ML.

```sql
CREATE OR REPLACE MODEL `your_project.your_dataset.churn_model` OPTIONS(model_type='logistic_reg')
AS SELECT feature1, feature2, ..., label
FROM `your_project.your_dataset.your_table`
 WHERE split_col = 'train';
```

### Step 4: Evaluate the Model

1. **Evaluate the model**: Use the evaluation dataset to check the performance of your model.

```sql
SELECT *
FROM
ML.EVALUATE(MODEL `your_project.your_dataset.churn_model`,
( SELECT feature1, feature2, ..., label
FROM `your_project.your_dataset.your_table` WHERE split_col = 'eval'))
```

### Step 5: Predict Churn

1. **Make predictions**: Use your trained model to predict churn on new or existing data.

```sql
SELECT customer_id, predicted_label, predicted_probability
FROM ML.PREDICT(MODEL `your_project.your_dataset.churn_model`,
( SELECT customer_id, feature1, feature2, ...
 FROM `your_project.your_dataset.new_data_table`))
```

### Step 6: Review and Iterate

1. **Review the results**: Analyze the prediction results and model performance. Look at metrics such as accuracy, precision, recall, and AUC-ROC.
2. **Iterate and improve**: Based on the evaluation results, you might want to refine your features, adjust the model type or parameters, and retrain the model

### Things to concider 🦧

In linear regression, the correlation matrix is a useful tool to identify which independent variables (features) have a strong linear relationship with the dependent variable (the variable you are trying to predict). Here’s how you can use the correlation matrix to decide which columns to consider:

### Steps to Use the Correlation Matrix for Feature Selection in Linear Regression

1. **Calculate the Correlation Matrix**:
   - The correlation matrix shows the correlation coefficients between all pairs of variables in your dataset. The values range from -1 to 1.
   - A value closer to 1 indicates a strong positive correlation, a value closer to -1 indicates a strong negative correlation, and a value around 0 indicates no correlation.
2. **Identify the Dependent Variable**:
   - Locate the row or column corresponding to the dependent variable in the correlation matrix.
3. **Select Relevant Features**:
   - Look at the correlation coefficients between the dependent variable and each of the independent variables.
   - Choose the independent variables that have a strong correlation with the dependent variable. There is no strict rule, but typically you might consider variables with an absolute correlation coefficient greater than 0.3 or 0.5, depending on your specific context and dataset.

### Example

Assume you have the following correlation matrix for a dataset with features `X1`, `X2`, `X3`, and a dependent variable `Y`:

| | X1 | X2 | X3 | Y |
|---|---|---|---|---|
| X1 | 1 | 0.6 | 0.1 | 0.8 |
| X2 | 0.6 | 1 | 0.3 | 0.7 |
| X3 | 0.1 | 0.3 | 1 | 0.2 |
| Y | 0.8 | 0.7 | 0.2 | 1 |

- **Dependent Variable (`Y`)**: The column and row corresponding to `Y` contain the correlation coefficients between `Y` and the features `X1`, `X2`, and `X3`.
- **Features to Consider**:
  - `X1` has a correlation of 0.8 with `Y`.
  - `X2` has a correlation of 0.7 with `Y`.
  - `X3` has a correlation of 0.2 with `Y`.

Based on these values, you would likely select `X1` and `X2` as the independent variables for your linear regression model because they have strong correlations with `Y`. `X3` has a weak correlation with `Y` and might not be as useful.

### Additional Considerations

- **Multicollinearity**: Check for multicollinearity among the independent variables. If two features are highly correlated with each other, it might cause issues in the regression model. You can use Variance Inflation Factor (VIF) to assess multicollinearity.
- **Domain Knowledge**: Use domain knowledge to ensure that the selected features make sense in the context of the problem you are solving.
- **Model Performance**: After selecting features based on the correlation matrix, it's essential to evaluate the model's performance. Sometimes, features that don't have a strong linear correlation might still be useful in the presence of other features.

By following these steps, you can use the correlation matrix effectively to select relevant features for your linear regression model.