PREDICTIVE MODEL PLAYBOOK

LINEAR REGRESSION

Summary: Linear Regression is a statistical method used to predict numeric outcomes by analyzing the outcome's relationship with one or more predictor variables.

STEP 1: SELECT TARGET AND PREDICTOR VARIABLES

Target Variable: The target variable is the variable we are trying to predict with the model. This should be continuous numeric variable, such as price, revenue, customers, etc.

Predictor variables: The predictor variables are used to help predict the target variable. Predictor variables should be: (1) Relevant to the target variable, (2) not highly correlated to other predictor variables, and (3), do not have a high number of missing values

Association Analysis

STEP 2: PREPARE DATA

Preparing the data includes dealing with issues such as missing, dirty, or duplicate data; removing outliers; blending and formatting data, etc. Your final dataset should include one row for each outcome and set of predictor variables.

Estimation and validation samples: Next, split the data set into two parts: one part for Estimation (for training the model) and one part for Validation (to help us verify that we are creating a useful model).

Create Samples

STEP 3: BUILD AND RUN THE MODEL

Run the model with the target and predictor variables. Observe the statistical significance of each of the predictor variables by looking at the p-value in the output. If it's below 0.05, then the relationship between the target and predictor variable is statistically significant. If not, it is not significant and can be excluded from the model. R-squared is an estimate between 0 and 1 of the explanatory power of them model, and can be used to compare models and select the best one.

Using a technique called "stepwise regression" can automatically identify the best combination of predictor variables.

Linear Regression

Stepwise

STEP 4: MODEL VALIDATION

Apply the model to the validation sample and observe how accurately the model predicts the outcomes. This step helps avoid overfitting and helps you understand how accurate your predictions will be on new data.

Model Comparison

STEP 5: APPLY THE MODEL TO MAKE PREDICTIONS

Apply the model to a new dataset to make predictions. This dataset should have all the predictor variable values, which are passed through the model to predict the unknown target variable value.

Score