**Analysis Planning Worksheet**

How did Zillow go so wrong in its machine learning algorithm, and how can we do better?

Zillow had a short-lived home-buying division to purchase homes, modestly renovate them and resell them at a profit. They even kicked off with a machine learning competition with a $1 million prize going to the team who could create an algorithm that would predict housing values. Zillow was forced to shutter that effort after just 9 months, when an unexpected dip in the market cost them nearly $500 million and they were left holding some 7,000 homes.

**Evaluation Question**

**Independent Variable(s)**

These variable(s) are causing something or creating an effect. List what each is and whether it is categorical or continuous. It is ok to only have one.

**Variable**

There are 79 independent variables; the categorical variables have numbers of levels. The continuous predictor variables are integers or floats.

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

**Variable**

□ Categorical : # of levels \_\_\_\_\_ □ Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Variable

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Dependent Variable(s)

These variable(s) are influenced by your independent variable and *depend* on them. List what each is and whether it is categorical or continuous. Unless they are related, you should have only one.

**Variable**

Housing price

□ Categorical: # of levels \_\_\_\_\_ X□ Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Variable

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Now that you know the type and number of independent and dependent variables, you are ready to use the analysis flow charts to choose your analysis!

**Analysis:**

Pearson Correlation, Exploratory Factor Analysis, Multiple Linear Regression and Stepwise Linear Regression for the continuous IV. ANCOVA or ANOVA for the categorical IVs.

**Analysis Planning Worksheet**

**Evaluation Question**

Which independent variables are the most important in predicting housing prices and which are just noise?

**Independent Variable(s)**

These variable(s) are causing something or creating an effect. List what each is and whether it is categorical or continuous. It is ok to only have one.

**Variable**

There are 79 independent variables – both categorical and continuous.

□ Categorical: # of levels \_\_varies □ Continuous

**Variable**

□ Categorical : # of levels \_\_\_\_\_ □ Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Variable

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Dependent Variable(s)

These variable(s) are influenced by your independent variable and *depend* on them. List what each is and whether it is categorical or continuous. Unless they are related, you should have only one.

**Variable**

Housing price

□ Categorical: # of levels \_\_\_\_\_ X Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Variable

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Now that you know the type and number of independent and dependent variables, you are ready to use the analysis flow charts to choose your analysis!

**Analysis:**

Use Correlation and stepwise multiple linear regression for the continuous ind variables. Check for mediation and moderation while running stepwise LR – other variables could influence the dep variable.

For the categorical IV, independent t test, ANCOVA and ANOVA.

**Analysis Planning Worksheet**

**Evaluation Question**

In the absence of having a human monitor the accuracy of a machine learning model, are there any variables I can include that could alert the algorithm that conditions are changing at the earliest opportunity?

**Independent Variable(s)**

These variable(s) are causing something or creating an effect. List what each is and whether it is categorical or continuous. It is ok to only have one.

**Variable**

Unemployment numbers from the Bureau of Labor Statistics

□ Categorical: # of levels \_\_\_\_\_ X Continuous

**Variable**

Defaults on Loans and Home mortgages

□ Categorical : # of levels \_2\_\_\_ □ Continuous

**Variable**

The price of lumber

□ Categorical: # of levels \_\_\_\_\_ X Continuous

Variable

Variations in Covid levels

□ Categorical: # of levels \_\_\_\_\_ X Continuous

Dependent Variable(s)

These variable(s) are influenced by your independent variable and *depend* on them. List what each is and whether it is categorical or continuous. Unless they are related, you should have only one.

**Variable**

Housing price

□ Categorical: # of levels \_\_\_\_\_ X Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

**Variable**

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Variable

□ Categorical: # of levels \_\_\_\_\_ □ Continuous

Now that you know the type and number of independent and dependent variables, you are ready to use the analysis flow charts to choose your analysis!

**Analysis:**

Structural equation modeling, exploratory factor analysis, correlations – pearson or Spearman rank.