Predict Audience count(#Audience) and Number of play count(#PlayCount):

We have a very rich dataset in content and features from Echonest. They have also released a [Taste Profile dataset](http://labrosa.ee.columbia.edu/millionsong/tasteprofile). This profile dataset has 1 million users and their listening data of 1 million MSD songs. We combined this dataset with the original Million Song Dataset.

Prediction of possible audience count and Number of play count is a regression problem. First step towards achieving this is preprocessing this dataset and cleaning the data.

**Data Preparation:**

The data is more or less preprocessed already but we ensured its cleaning by cleaning the missing data and removing duplicate rows.

Azure ML Studio picked few numeric features as string so we had to convert them to numeric features by using Edit Metadata block.

Some other preprocessing techniques that we ensured are

1. Removing all nan values.
2. Converting the “Year” attribute into categorical that would map each year to that decade.

e.g. All years from 1971 to 1980 are converted to 1980. This converted all years from 1922 to 2010 to their respective decades and resulted in 9 unique categories.

1. Artist has Music Brainz tags like pop, rock and indie etc. We found out the most common tags for all the artist and mapped these tags to only 12 categories. So 11 unique categories and all others are mapped as a single category “Others”.
2. Removed all columns that are less of importance with respect to Target variable. (#Audience and #PlayCount)

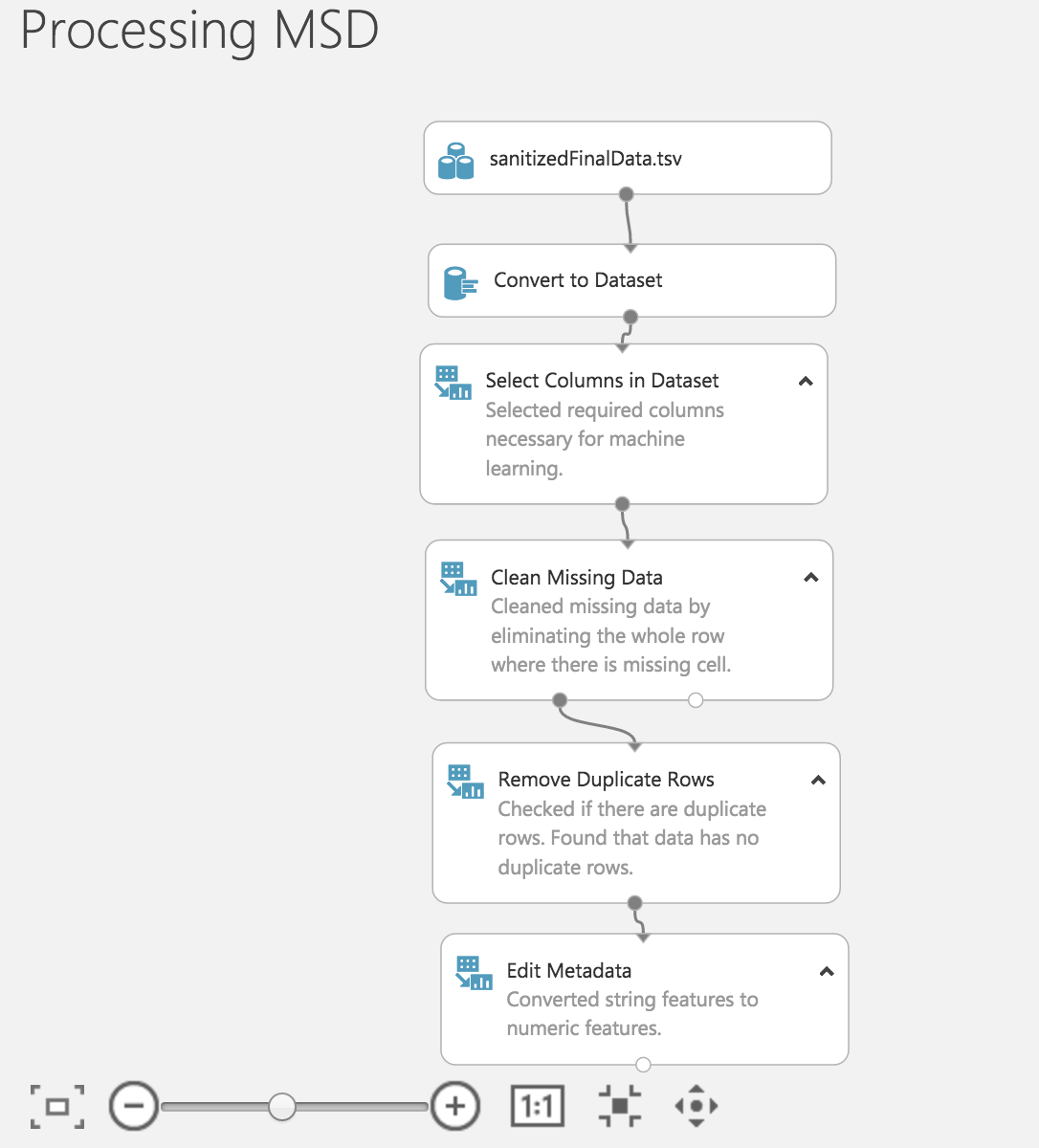


Figure 1: Data Preparation

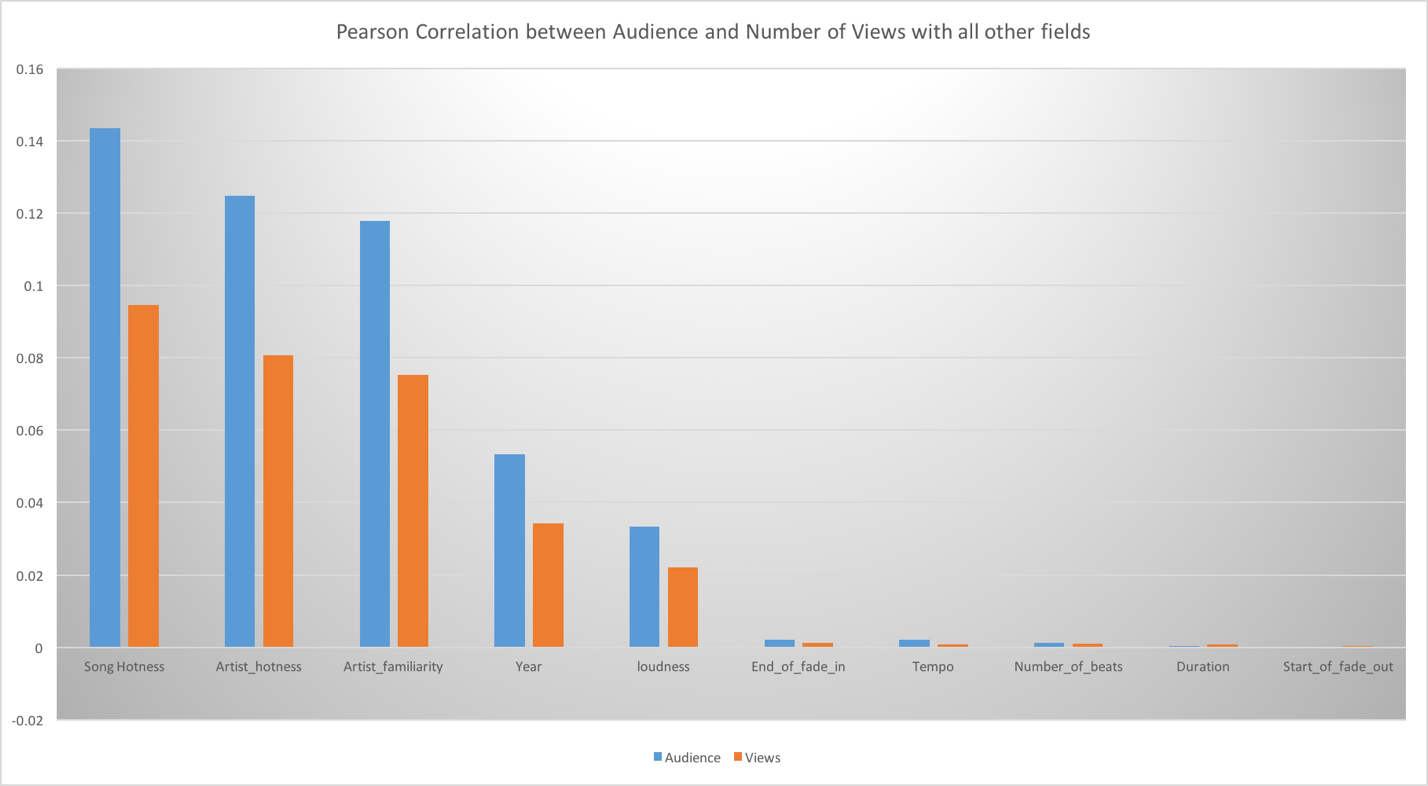


Figure 2 Pearson Correlation Coefficient with respect to Targets

1. Pearson Correlation Coefficient is calculated between Target variables and all other independent attributes of the MSD dataset. This graph shows that “Song Hotness” is positively correlated with Audience and Views. This gives good indication of number of columns that will go as input in Regression Model.

**Model:**

1. We split the dataset in 80:20 proportion between training and test set. This did not give the results as accurate as the cross validation model.
2. The model bias is eliminated using Cross Validation module to ensure that the model is trained on all possible

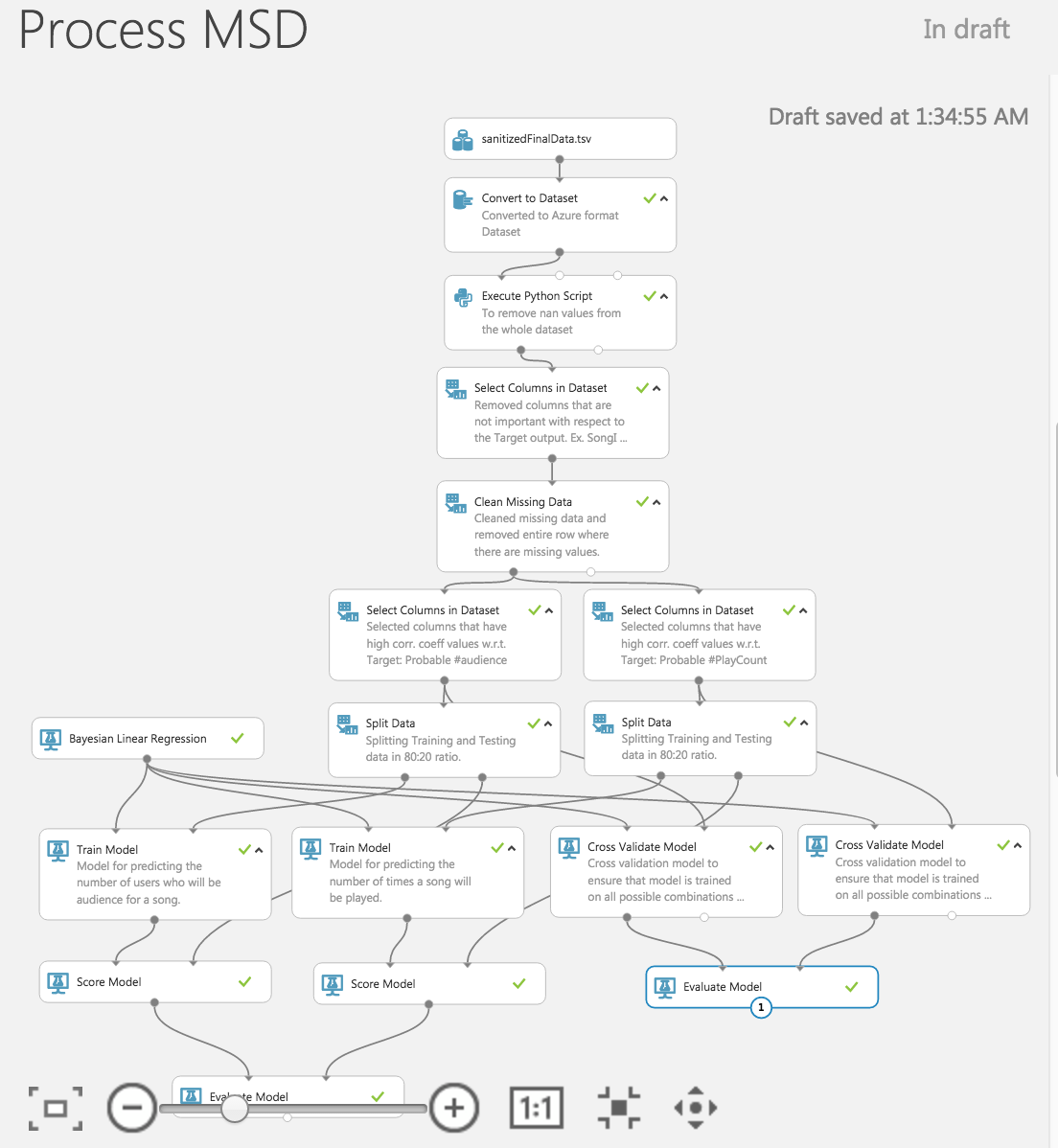


Figure 3 Linear Regression Model

