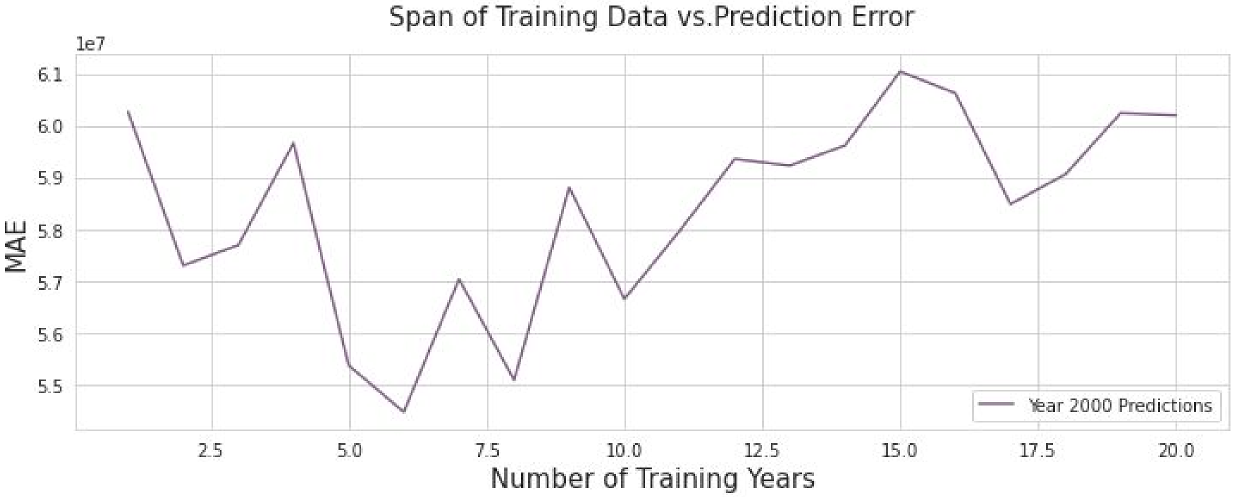
Capstone 2: Predictions

After the best model was found, some predictions were made on a subset of the dataset in the Predict by Year notebook. The year 2000 was chosen as the target year for predicting. In order to find the optimum range of years to select for the training data, the prediction scores for many ranges were tried. It was assumed that the most recent years, prior to the year 2000, would be the best ones to use for training. The widest range of training years tried was 20 years.



Using a span of 6 years prior to the year 2000 produced the lowest prediction error. The training samples were 2433 in number, and the prediction labels had a count of 831. An average of 864 boosting rounds were used. The 10 fold cross validation (CV) mean absolute error (MAE) on the holdout set from the training data was $32,211,893. This was very similar to what was observed when the final model was validated, after the Hyperparameter tuning.

Using this span of training years to make revenue predictions for actors who appeared in films in the year 2000, with the best boosted model, gave a final CV prediction MAE of $54,470,586. The smaller training sample size and the fact that the best model was very complex contributed to this inflated error, compared with the CV test error on the training data.