Mike Rocchio

MSDS 411 Project Description

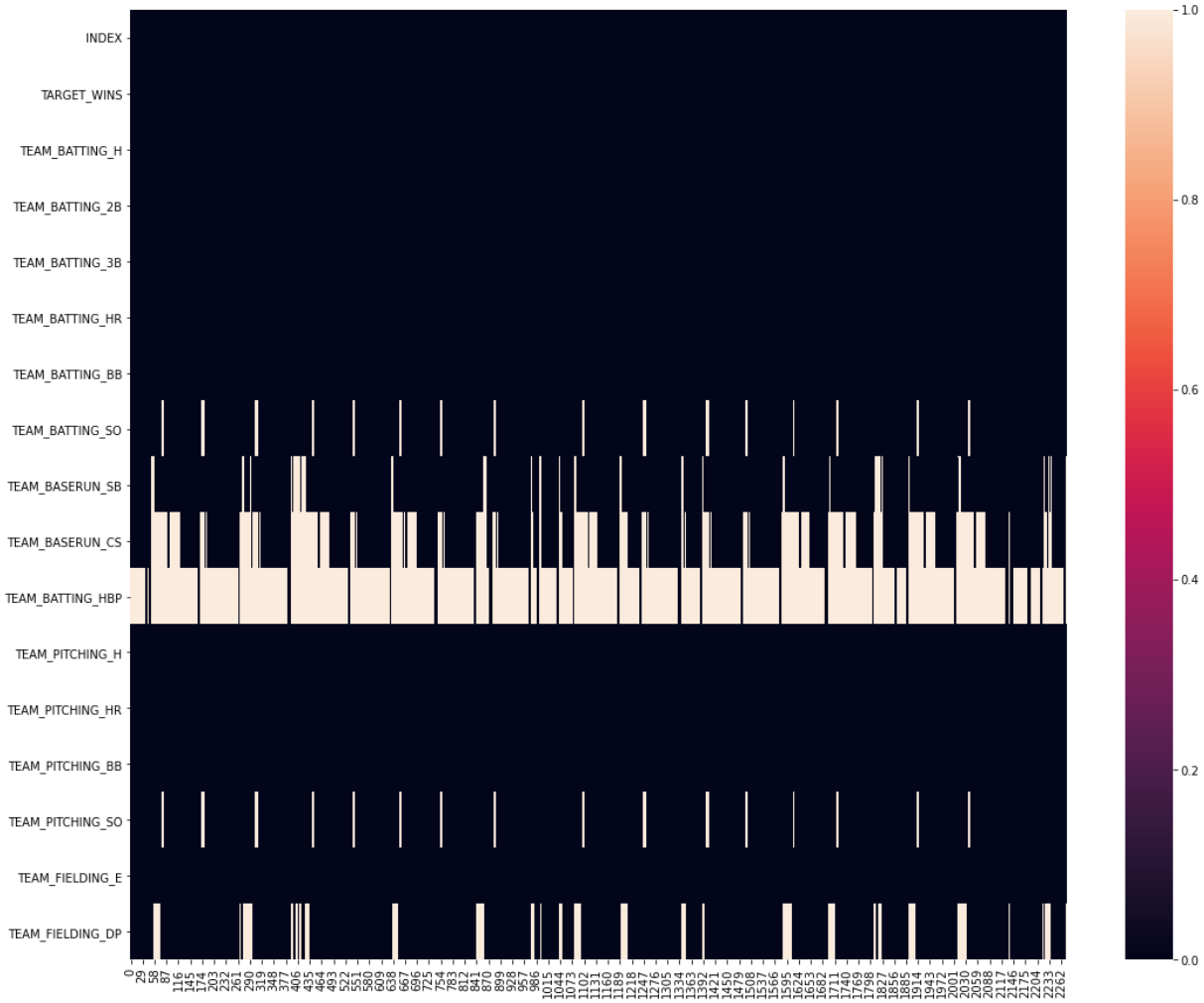
08/11/2021

Note on paper length: It was stated that the target audience is a product/project manager and we should be giving them a summary of our model. I have never in my life seen or heard of an executive summary over 5 pages for a single model, anything long just won't get read. This is what I would typically write for a PM I am working to develop a model with.

Executive Summary for Project Management Staff – Model #1

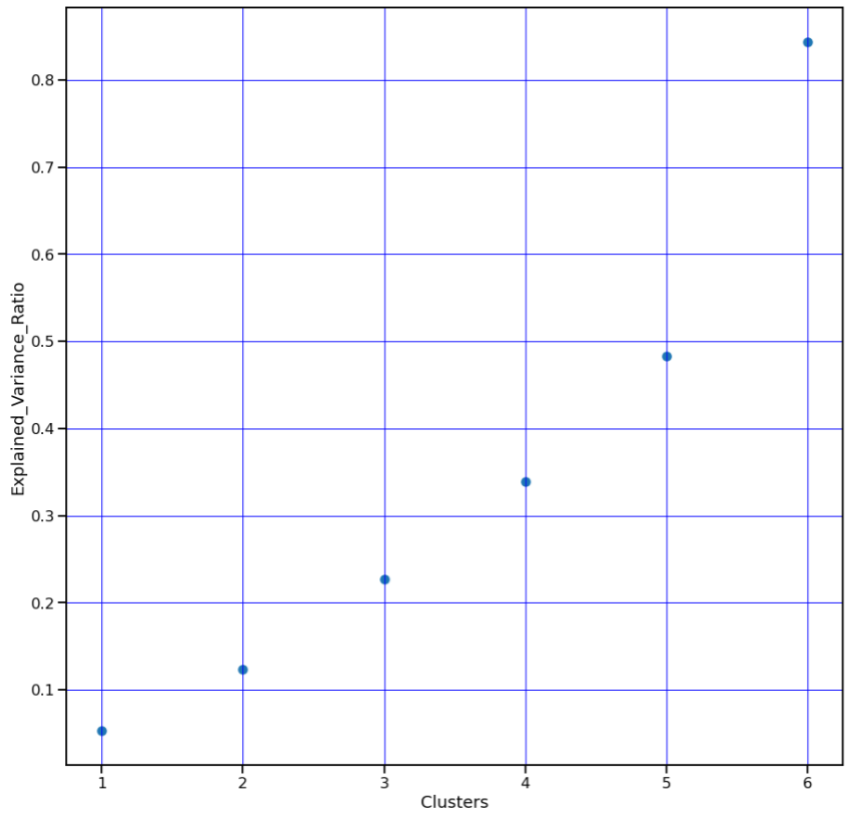
I. Overview of the Dataset

There was one primary issue within this dataset that could contribute to production model risk. There were several null fields (missing values) and we determined that they are most likely representing values that were calculated to be zero. However, we didn't have any data governance to reach out to in order to resolve this issue and it may increase the risk of the end model. Below is a visual of those fields with white being the missing values for each column:

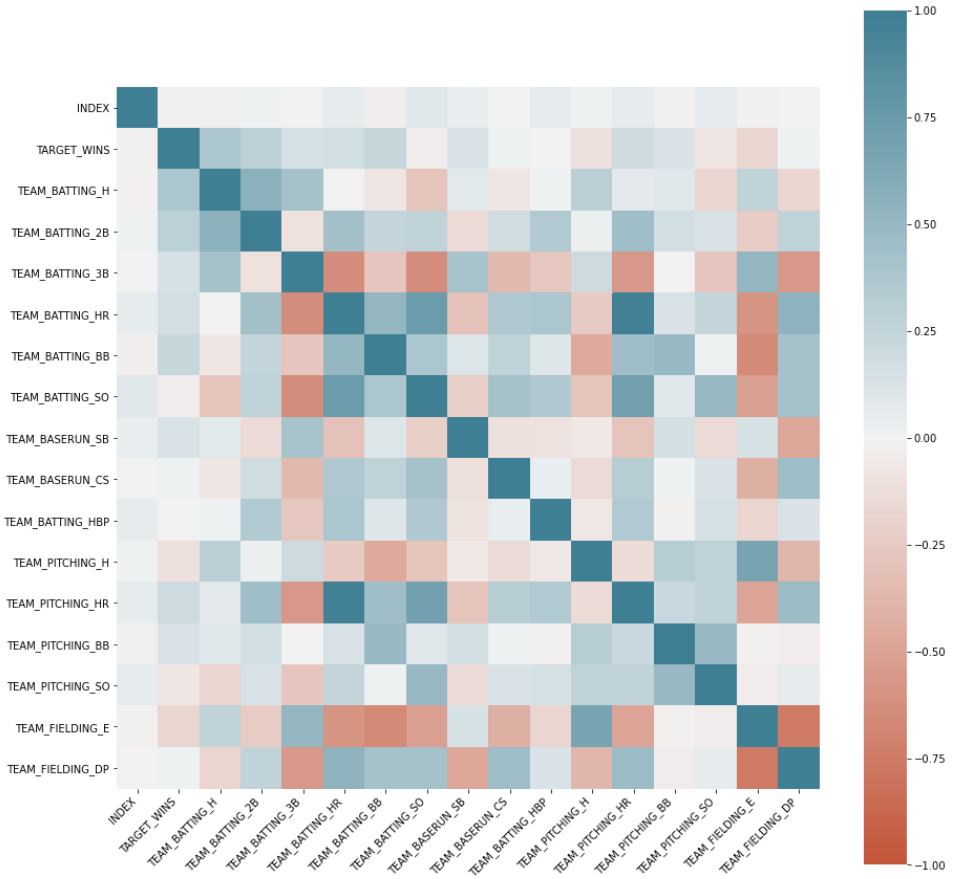


II. Overview of the methods used to train the model

Step 1 PCA Analysis - Due to the high dimensionality and the limited number of records within the dataset we decided to first conduct a principal component analysis. Simplified, this takes all the fields within the data and compresses them into a few fields. We typically want to see, within those newly created fields (called principal components), a representation of at least 80% of the variance within the dataset. To reach this threshold we increase the number of components gradually. For this analysis we utilized 6 components as this was determined to be the amount at which the threshold is reached:



Step 2 – Simple Linear Regression – After we calculated these six components, we elected to conduct a linear regression of the newly created principal components. This method is simple and easy and creates an easy to maintain production quality model. This method also contains a low level of model risk when introduced into a production environment. Below is a visualization that speaks to the high dimensionality of the data and why a linear regression could not be initially conducted, please note how scattered the values within the correlation matrix(blue denotes a positive correlation, white denotes no correlation, and red denotes an inverse correlation:



III. **Results & Model Risk Assessment**

Results – This model had a very high r2_score and a very low MSE, meaning that it is very accurate. Below are some selected outputs and visualizations from our best training run:

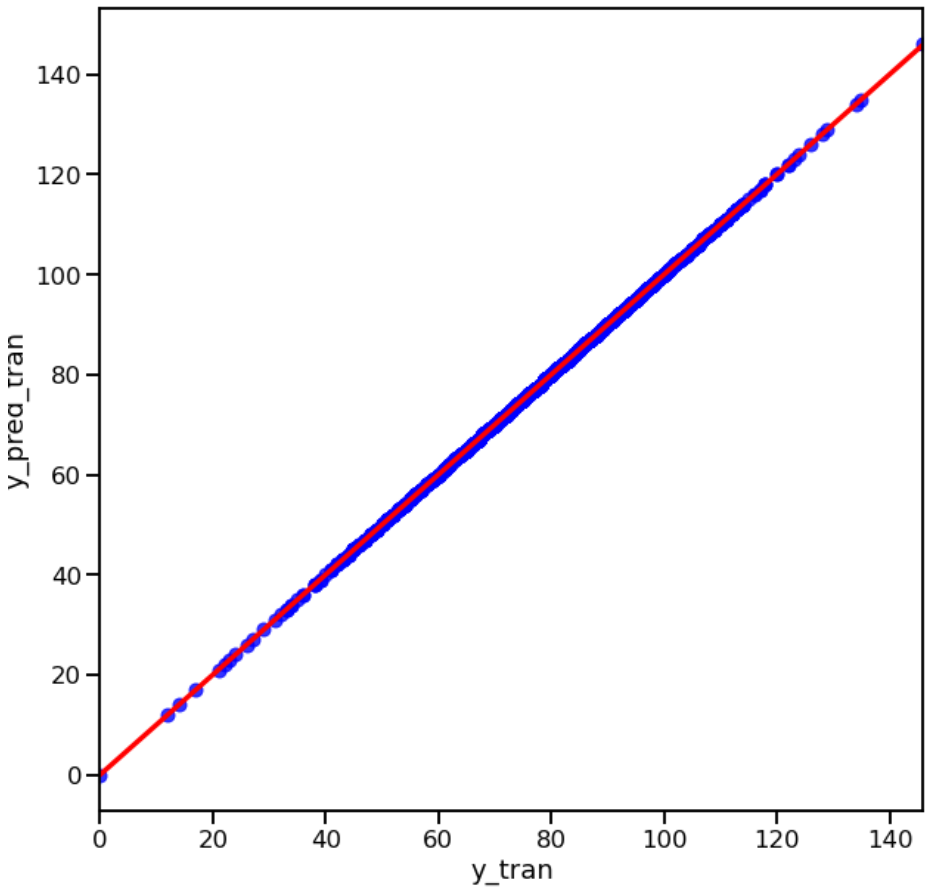
r2_score: 0.98

r2_cross_val: 0.98

mse: 1.189 * e-32

mse_cross_val : 5.348 * e-30

***Note**: For details on these metrics please refer to our SharePoint prior to placing ticket or setting a follow up with our analytics team.*



Model Risk Assessment, High Level Overlay – Prior to model testing and QA assessment our assessment is that this model contains one primary risk caused by the null fields of the dataset. However, we also assessed that as L1 and therefore negligible upon initial review.

Please reach out to us if you have any questions or concerns.