

n [88]:	<pre>test_df[test_df.RATING == "A0"] ID CONSOLE YEAR CATEGORY</pre>
	train_df.head() ID CONSOLE YEAR CATEGORY PUBLISHER RATING CRITICS_POINTS USER_POINTS SalesInMillions 0 2860 ds 2008 role-playing Nintendo E 2.833333 0.303704 0.623434 1 731 wii 2012 simulation Konami Digital Entertainment E10+ 13.200000 1.640000 -1.257900 2 495 pc 2019 shooter Activision M 4.562500 0.006410 -0.576469 3 2641 ps2 2002 sports Electronic Arts E 4.181818 0.326923 0.339633
in [90]: out[90]:	4 811 ps3 2013 action Activision M 2.259259 0.032579 -1.814317 combine.skew() ID -0.025973 YEAR 0.060762 CRITICS_POINTS 2.299211 USER_POINTS 1.607630 dtype: float64
n [91]:	Reduced skewness of CRITICS POINTS by applying boxcox transformation modify_a_col(combine.CRITICS_POINTS, -0.3) train_df["CRITICS_POINTS"] = boxcox(train_df["CRITICS_POINTS"], -0.3) test_df["CRITICS_POINTS"] = boxcox(test_df["CRITICS_POINTS"], -0.3) 0.030325075809584566
	0.030325075809584566 0.8 0.7 0.6 0.5 0.4 0.3
	Reduced skewness of USER POINTS by applying boxcox transformation
n [92]:	<pre>modify_a_col(combine.USER_POINTS, 0.26) train_df["USER_POINTS"] = boxcox(train_df["USER_POINTS"], 0.26) test_df["USER_POINTS"] = boxcox(test_df["USER_POINTS"], 0.26) 0.024270050203997705 0.40</pre>
	0.35 - 0.30 - 0.25 - 0.20 - 0.15 - 0.10 - 0.05 -
	0.00
	<pre>"E" : 1, "E10+" : 5, "M" : 10, "T" : 15 } # train_df["RATING"] = train_df["RATING"].map(rating_dict) # test_df["RATING"] = test_df["RATING"].map(rating_dict) # combine["RATING"] = combine["RATING"].map(rating_dict)</pre>
n [97]:	<pre>I tried Feature Engineering but that decreased model accuracy # train_df["Skills"] = inv_boxcox(train_df.USER_POINTS, 0.26) * train_df.RATING # test_df["Skills"] = inv_boxcox(test_df.USER_POINTS, 0.26) * test_df.RATING # test_df["Skills"] = test_df["Skills"].fillna(np.mean(train_df.Skills)) # combine["Skills"] = inv_boxcox(combine.USER_POINTS, 0.26) * combine.RATING</pre>
n [98]: Out[98]:	<pre># combine["Skills"] = combine["Skills"].fillna(np.mean(train_df.Skills)) # train_df["Skills"] = boxcox(train_df.Skills, 0.20) # test_df["Skills"] = boxcox(test_df.Skills, 0.20) # combine["Skills"] = boxcox(combine.Skills, 0.20) train_df.skew() ID</pre>
	<pre>CRITICS_POINTS 0.018155 USER_POINTS 0.018782 SalesInMillions -0.001442 dtype: float64 # modify_a_col(train_df.Skills, 0.20) print(test_df.CONSOLE.unique()) ['ps2' 'psp' 'x' '3ds' 'ds' 'pc' 'x360' 'xone' 'wii' 'gc' 'psv' 'gba'</pre>
it[101]:	<pre>'ps3' 'wiiu' 'ps4' 'ps' 'dc'] len(train_df.CONSOLE.unique()) 17 print(train_df.CONSOLE.unique()) ['ds' 'wii' 'pc' 'ps2' 'ps3' 'psp' 'x' 'x360' 'psv' 'gba' 'gc' 'ps' 'wiiu' 'xone' 'ps4' '3ds' 'dc']</pre>
n [65]:	"x" : "x-type", "xone" : "x-type", "x360" : "x-type", "psp" : "p-type", "ps2" : "p-type", "ps3" : "p-type", "ps4" : "p-type", "ps" : "p-type",
	<pre>"psv" : "p-type", "pc" : "pc", "wii" : "w-type", "ds" : "d-type", "3ds" : "d-type", "dc" : "d-type", "gba" : "g-type", "gc" : "g-type", }</pre>
	<pre>print(len(console_dict.keys())) print(set(console_dict.keys()) == set(train_df.CONSOLE.unique())) # train_df["CONSOLE"] = train_df.CONSOLE.map(console_dict) # test_df["CONSOLE"] = test_df.CONSOLE.map(console_dict) # combine["CONSOLE"] = combine.CONSOLE.map(console_dict)</pre> 17 True
	<pre># train_df["How_much_Old"] = train_df["YEAR"].apply(lambda x : 2020 - x) # test_df["How_much_Old"] = test_df["YEAR"].apply(lambda x : 2020 - x) # combine["How_much_Old"] = combine["YEAR"].apply(lambda x : 2020 - x) # train_df = train_df.drop("YEAR",axis = 1) # test_df = test_df.drop("YEAR",axis = 1) # combine = combine.drop("YEAR",axis = 1)</pre> # def move_target_to_last(dataset, col_name): <pre></pre>
	<pre>x = dataset[col_name] dataset = dataset.drop(col_name, axis = 1) dataset[col_name] = x return dataset train_df = move_target_to_last(train_df, "SalesInMillions")</pre>
[105]: t[105]:	ID CONSOLE YEAR CATEGORY PUBLISHER RATING CRITICS_POINTS USER_POINTS SalesInMillions
[106]: t[106]:	4 811 ps3 2013 action Activision M 0.723042 -2.267121 -1.814317 test_df.head() ID CONSOLE YEAR CATEGORY PUBLISHER RATING CRITICS_POINTS USER_POINTS 0 4310 ps2 2008 action Tecmo Koei T 1.122234 -0.664144 1 4011 psp 2007 strategy Atari E10+ 1.338705 -1.907006 2 2185 ps2 2004 shooter Electronic Arts T 0.944129 -1.979223
[107]: t[107]:	3 1644 x 2006 action Electronic Arts E 0.914674 -0.125732 4 188 3ds 2011 racing Ubisoft E10+ 0.146945 -1.371757 train_df.corr() ID YEAR CRITICS_POINTS USER_POINTS SalesInMillions ID 1.0000000 -0.052969 0.007047 0.061334 -0.032628
[108]:	YEAR -0.052969 1.000000 -0.181344 -0.415385 0.004861 CRITICS_POINTS 0.007047 -0.181344 1.000000 0.521520 -0.119885 USER_POINTS 0.061334 -0.415385 0.521520 1.000000 -0.205847 SalesInMillions -0.032628 0.004861 -0.119885 -0.205847 1.000000 new_train = train_df.copy() new_test = test_df.copy()
[109]: t[109]:	new_train = new_train.drop("CRITICS_POINTS", axis = 1) new_test = new_test.drop("CRITICS_POINTS", axis = 1) train_df.head() ID CONSOLE YEAR CATEGORY PUBLISHER RATING CRITICS_POINTS USER_POINTS SalesInMillions 0 2860 ds 2008 role-playing Nintendo E 0.894459 -1.024760 0.623434 1 731 wii 2012 simulation Konami Digital Entertainment E10+ 1.796218 0.527919 -1.257900
ո [110]։	<pre>2 495 pc 2019 shooter</pre>
	<pre>return False def evaluate(self, approxes, target, weight): assert len(approxes) == 1 assert len(target) == len(approxes[0]) approx = approxes[0] error_sum = 0.0 weight_sum = 0.0</pre>
	<pre>for i in range(len(approx)): w = 1.0 if weight is None else weight[i] weight_sum += w error_sum += inv_boxcox(w,0.27) * ((inv_boxcox(approx[i],0.27) - inv_boxcox(target[i],0.27))**2) return error_sum, weight_sum Since we have transformed output variable we need to change eval metric so that catboost model performance increase</pre>
[111]:	<pre>Model score increased from 1.66 to 1.62 without any tuning # class RmseObjective(object): # def calc_ders_range(self, approxes, targets, weights): # assert len(approxes) == len(targets) # if weights is not None: # assert len(weights) == len(approxes) # targets = inv_boxcox(targets, 0.27)</pre>
	<pre># approxes = inv_boxcox(approxes, 0.27) # result = [] # for index in range(len(targets)):</pre>
[112]:	<pre># der2 *= weights[index] # result.append((der1, der2)) # return result Getting Validation Test X_train, X_test, y_train, y_test = train_test_split(train_df.drop(["SalesInMillions"], axis = 1), t in_df["SalesInMillions"], test_size = 0.2, random_state=8)</pre>
	<pre>cat_columns=train_df.select_dtypes(include=['object', 'category']).columns.tolist() cat_columns ['CONSOLE', 'CATEGORY', 'PUBLISHER', 'RATING'] Model Training</pre>
[114]:	<pre>model = CatBoostRegressor(use_best_model=True, learning_rate=0.01, n_estimators=10000, objective="RMSI" , cat_features=cat_columns, eval_metric=Jiten())</pre>















































