Model	Parameters	Class	Training - all features				Test - all features				Training - 10 features				Test - 10 features			
			Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
Naïve-Bayes (Gaussian)		0	0.7	0.82	0.97	0.72	0.69	0.82	0.97	0.72	0.81	0.9	0.94	0.86	0.81	0.9	0.94	0.86
		1		0.21	0.16	0.29		0.23	0.18	0.32		0.14	0.15	0.13		0.14	0.15	0.13
		2		0.36	0.23	0.78		0.35	0.23	0.78		0.49	0.37	0.74		0.5	0.37	0.75
		3		0.08	0.05	0.19		0.09	0.05	0.21		0.06	0.04	0.11		0.04	0.03	0.08
Naïve-Bayes (Categorical)		0	0.93	0.97	0.97	0.96	0.91	0.95	0.97	0.94	0.88	0.94	0.93	0.95	0.85	0.93	0.92	0.94
		1		0.71	0.75	0.66		0.52	0.59	0.46		0.22	0.66	0.13		0.04	0.11	0.03
		2		0.78	0.69	0.89		0.72	0.62	0.85		0.52	0.44	0.63		0.43	0.37	0.52
		3		0.55	0.47	0.66		0.39	0.3	0.57		0.17	0.69	0.1		0.02	0.11	0.01
Logistic Regression [album, non-album]	C=0.1, penalty='l2'	0	0.904	0.95	0.92	0.97	0.903	0.95	0.92	0.95	0.887	0.94	0.92	0.96	0.886	0.94	0.92	0.95
		1		0.53	0.7	0.43		0.53	0.69	0.53		0.48	0.57	0.41		0.48	0.57	0.41
SVM Linear Binary [album, non-album]	C=100, random_state=42, max_iter=25000, loss=squared_hinge	0	0.87	0.92	0.95	0.9	0.87	0.92	0.95	0.9	0.89	0.94	0.93	0.94	0.89	0.94	0.93	0.94
		1		0.55	0.49	0.64		0.55	0.48	0.64		0.53	0.57	0.5		0.53	0.55	0.5
SVM Linear	C= 0.01,random_state=42, max_iter = 3000,loss="squared_hinge"	0	0.85	0.93	0.89	0.97	0.85	0.93	0.89	0.96	0.84	0.92	0.88	0.96	0.84	0.91	0.88	0.95
		1		0.12	0.13	0.11		0.11	0.11	0.1		0.01	0.25	0		0	0.12	0
		2		0.13	0.58	0.08		0.14	0.6	0.08		0.14	0.49	0.08		0.15	0.51	0.09
		3		0	0	0		0	0	0		0.02	0.01	0.04		0.02	0.01	0.05
SVM NonLinear	kernel = rbf, gamma= auto	0	0.92	0.96	0.94	0.99	0.91	0.96	0.93	0.98	0.88	0.94	0.91	0.97	0.88	0.94	0.91	0.97
		1		0.28	0.82	0.17		0.26	0.77	0.16		0.04	0.61	0.02		0.04	0.59	0.02
		2		0.67	0.67	0.67		0.65	0.65	0.65		0.48	0.48	0.48		0.48	0.47	0.48
		3		0.58	0.97	0.42		0.5	0.95	0.34		0	0	0		0	0	0
SVM NonLinear	kernel = polynomial, gamma= auto	0	0.92	0.96	0.94	0.98	0.92	0.96	0.94	0.98	0.88	0.94	0.91	0.97	0.88	0.94	0.91	0.97
		1		0.32	0.83	0.2		0.3	0.77	0.19		0.02	0.61	0.01		0.02	0.77	0.01
		2		0.67	0.67	0.67		0.66	0.66	0.66		0.49	0.47	0.52		0.49	0.47	0.52
		3		0.67	0.97	0.51		0.55	0.88	0.4		0.02	1	0.01		0.01	1	0.01
Rule-Based Classifier	k=1, prune_size=0.33	0	0.922	0.95	0.92	1	0.921	0.93	0.93	0.99	0.874	0.93	0.88	1	0.873	0.93	0.88	0.99
		1		0.44	0.94	0.29		0.41	0.84	0.28		0.01	0.62	0.01		0.01	0.43	0.01
		2		0.64	0.91	0.5		0.66	0.88	0.53		0.04	0.5	0.02		0.104	0.46	0.06
		3		0.57	0.94	0.41		0.52	0.8	0.39		0.01	0.29	0.01		0.01	0.07	0.01
EC Random Forest	n_estimators=100, criterion="gini", max_depth=17, min_samples_split=3, min_samples_leaf=3, max_features=" auto", random_state=10, class_weight=" balanced"	0	0.95	0.97	1	0.95	0.94	**0.97**	0.99	0.95	0.82	0.9	0.98	0.84	0.78	0.88	0.95	0.82
		1		0.81	0.72	0.92		**0.74**	0.66	0.84		0.35	0.27	0.51		0.17	0.13	0.24
		2		0.83	0.72	0.97		**0.79**	0.69	0.93		0.59	0.45	0.86		0.52	0.39	0.76
		3		0.83	0.72	0.99		**0.7**	0.66	0.75		0.41	0.27	0.92		0.16	0.11	0.33
EC-Bagging (Decision Tree)	criterion="gini", max_depth=9, min_samples_split=10, min_samples_leaf=10	0	0.93	0.96	0.94	0.99	0.92	0.96	0.94	0.99	0.89	0.94	0.91	0.98	0.88	0.94	0.91	0.97
		1		0.42	0.8	0.28		0.43	0.78	0.3		0.07	0.74	0.04		0.06	0.59	0.03
		2		0.71	0.74	0.69		0.69	0.71	0.68		0.48	0.51	0.45		0.45	0.47	0.43
		3		0.44	1	0.28		0.4	1	0.25		0.26	0.89	0.15		0.21	0.88	0.12
EC-Boosting (Decision Tree)	criterion="gini", max_depth=9, min_samples_split=10, min_samples_leaf=10	0	0.98	0.99	0.99	0.99	0.94	0.97	0.96	0.99	0.89	0.95	0.95	0.95	0.83	0.92	0.91	0.93
		1		0.88	0.91	0.85		0.65	0.76	0.57		0.45	0.43	0.47		0.12	0.13	0.12
		2		0.9	0.91	0.89		0.75	0.8	0.7		0.55	0.56	0.54		0.34	0.37	0.32
		3		0.98	0.98	0.97		0.6	0.92	0.44		0.75	0.78	0.72		0.15	0.28	0.1
EC-Bagging (Random Forest)	n_estimators=100, criterion="gini", max_depth=17, min_samples_split=3, min_samples_leaf=3, max_features=" auto", random_state=10, class_weight=" balanced"	0	0.95	0.97	0.99	0.95	0.94	**0.97**	0.99	0.95	0.84	0.92	0.96	0.87	0.82	0.91	0.95	0.87
		1		0.79	0.72	0.87		**0.74**	0.68	0.8		0.34	0.3	0.39		0.19	0.17	0.21
		2		0.8	0.69	0.95		**0.77**	0.66	0.93		0.57	0.44	0.83		0.52	0.4	0.76
		3		0.83	0.76	0.91		**0.67**	0.69	0.66		0.43	0.34	0.58		0.22	0.18	0.28
EC-Boosting (Random Forest)	n_estimators=100, criterion="gini", max_depth=17, min_samples_split=3, min_samples_leaf=3, max_features=" auto", random_state=10, class_weight=" balanced"	0	0.99	1	0.99	1	0.96	0.98	0.97	0.99	0.97	0.99	0.99	0.99	0.86	0.93	0.91	0.95
		1		0.96	0.99	0.94		0.78	0.94	0.66		0.9	0.91	0.89		0.14	0.2	0.11
		2		0.95	0.96	0.94		0.85	0.88	0.81		0.86	0.83	0.89		0.4	0.42	0.38
		3		1	1	1		0.69	0.96	0.54		0.99	0.99	1		0.23	0.54	0.14
Single hidden layer Neural Network	activation: identity learning_rate_inits: 0.02 hidden_layer_size: 40	0	0.89	0.95	0.91	0.99	0.91	0.96	0.94	0.98	0.87	0.94	0.92	0.96	0.88	0.94	0.92	0.96
		1		0.17	0.34	0.11		0.21	0.33	0.16		0.02	0.3	0.01		0.01	0.27	0
		2		0.51	0.67	0.41		0.56	0.58	0.55		0.48	0.42	0.54		0.46	0.4	0.55
		3		0	0	0		0.04	0.75	0.02		0	0	0		0	0	0
Single hidden layer Neural Network	activation: identity learning_rate_inits: 0.001 hidden_layer_size: 350	0	0.89	0.95	0.92	0.97	0.9	0.95	0.94	0.97	0.87	0.93	0.88	1	0.89	0.94	0.9	1
		1		0.13	0.38	0.08		0.19	0.44	0.12		0	0	0		0	0	0
		2		0.57	0.57	0.57		0.59	0.52	0.69		0.18	0.54	0.11		0.17	0.45	0.1
		3		0.01	0.03	0.01		0.04	0.26	0.02		0	0	0		0	0	0
Deep Neural Network	activation: identity learning_rate_inits: 0.001 hidden_layer_sizes: 40, 40	0	0.89	0.95	0.92	0.98	0.91	0.96	0.94	0.97	0.87	0.94	0.9	0.97	0.89	0.94	0.92	0.97
		1		0.15	0.35	0.1		0.2	0.38	0.13		0.01	0.28	0.01		0	1	0
		2		0.55	0.6	0.5		0.58	0.53	0.64		0.44	0.46	0.42		0.43	0.43	0.43
		3		0	0	0		0	0	0		0	0	0		0	0	0
Deep Neural Network	activation: identity learning_rate_inits: 0.001 hidden_layer_sizes: 40, 20, 8	0	0.89	0.95	0.92	0.98	0.91	0.96	0.93	0.98	0.87	0.93	0.87	1	0.89	0.94	0.89	1
		1		0.11	0.41	0.07		0.12	0.45	0.07		0	0	0		0	0	0
		2		0.56	0.61	0.51		0.59	0.55	0.62		0.09	0.47	0.05		0.06	0.34	0.03
		3		0	0	0		0	0	0		0	0	0		0	0	0

Figure 42: Main Results

6 Linear regression

6.1 Univariate problem

We decided to try and predict how many favorites a track will have based on how many listens it has. Therefore our regression problem is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where x is *listens*, y is *favorites*, the betas are **coefficients** and ϵ is the **residual**. We found excellent evaluation metrics considering that there's only one regressor, i.e. a very high **53%** R^2. (other metrics at the end)

However, we have to also recognize that this problem, as likely many other regression problems we could have chosen on this dataset, suffers from a **reverse causality** issue. Not only does x influence y, but as the number of favorites a track has does influence the **recommendation** algorithm and how often that track is suggested to users who haven't listened to it yet,

favorites also influence listens, that is y influences x. Therefore we cannot be sure that our regressor **correctly** captures the causal effect of x on y and the best we can do is underline the high covariance that these two variables capture of each other and avoid drawing other, stronger conclusions.

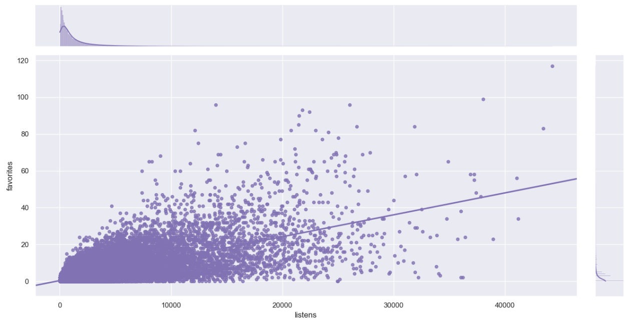

Figure 43: Linear univariate regression with scatterplot of datapoints and distribution curves

The graph shows a particularly **imbalanced distribution** of both of these variables, with one large high peak in the low values and relatively very few points in the top 2 quartiles. **OLS** regression is centered around the assumption of a **normal** distribution, which in this case, it's far from the truth. Therefore we should consider this as one of the problems warning us of the poor **predictive** capability of this regression function, notwithstanding the apparently good metrics.

6.2 Multivariate problem

The multivariate problem is the same effort of predicting *(track, favorites)* but using as regressors all continuous and binary variables instead of just one. This method and its results lead to some interesting considerations.

We got a **63.8 %** R^2 with a **7.600** MSE and **1.570** MAE.

First of all, this result is generally better in its assumed predictive capability than the univariate one, seeing as the metrics are better. However, we also have to consider two issues:

1. The addition of **24 regressors** only increased the R^2 by **20%** of the previous one. This also suggests an issue, just as much as it did that in the univariate problem only one variable could account for a 53% R^2.
2. The reverse causality issue in the univariate is getting bigger here in the multivariate. Anything that is done during music production could not be affected by how many favorites the track has after "commercial" release. However, we also have our three *comments* features that might be influenced by *favorites* in the same way that *listens* is: the recommendation algorithm suggests a song more as it has more favorites. However, this issue should be smaller in entity than that related to *listens*, as we believe the listens to be the main driver of the recommendations. Also, the reverse causality problem could be diluted as there are many features. Therefore, our **63.8%** R^2 could be closer to a genuine, issue-free R^2 than that of the univariate.

In conclusion, we would not use regression to predict any of these variables. There are too many unsolved doubts about distribution, reverse causality, possible omitted features and how the recommendation algorithm works. **OLS regression** requires strong assumptions in these fields and we're not confident we can make them with no negative repercussions. But, if forced to solve this problem with regression, I would choose the multivariate, in the hopes that our reverse causality could get **diluted**.

	All features			10 features		
Model	R^2	MSE	MAE	R^2	MSE	MAE
Multivariate Linear Regression	**63.8%**	**7.600**	**1.570**	57.7%	8.886	1.648
Multivariate Lasso	63.0%	7.774	1.528	57.7%	8.879	1.637
Multivariate Ridge	63.8%	7.600	1.570	57.7%	8.886	1.648
Univariate Linear Regression	**53.0%**	**10.510**	**1.673**			
Univariate Lasso	53.0%	10.510	1.673			
Univariate Ridge	53.0%	10.510	1.673			

Figure 44: Main Results