





#### **Dataset & Problem**

- "Spotify Sequential Skip Prediction Challenge" by Spotify on Alcrowd
- 2 datasets: "track features" and "log"
- Each dataset contains:
  - track ID, session ID, and other identifiers
  - musical information
  - user behavior and history
  - Skipped or not?

Challenge: Predict if a user will skip the current song



## **EDA - Heatmap of Correlation Coefficients**

session_position -	1	0.31	0.055	0.04	0.0012	0.0055	0.021	0.19	0.027	0.023	0.036	0.024	0.012	0.004	0.024	0.0048	0.00036	0.0047	0.00079	0.0012	0.0004	0.00028	0.0027	0.00071	0.002	0.0038	0.0035	0.0022	0.0026	0.0028	1.8e-06	0.0081	0.0022 -0	0039 0	0074 0	0083	0.0014	0.0026	0.0079	0.0028	.0082 0	0.0018 0	0038
session_length -	0.31	1	0.091					0.054																																			.0071
skip_1 -	0.055	0.091	1	0.82	0.63	-0.6																																					0.014
skip_2 -			0.82	1	0.76	-0.73																																					.0097
skip_3 -			0.63	0.76	1	-0.96																																					1.022
not_skipped -			0.6	0.73	0.96	1																																					0.016
context_switch -				0.074	0.043	0.036	1																																				0016
no_pause_before_play -						0.08	0.0039	1	-0.75	-0.83																																	.013
short_pause_before_play -							0.0097	-0.75	1	0.91																																	0.019
long_pause_before_play -							0.028	-0.83	0.91	1	0.0097																																0.014
hist_user_behavior_n_seekfwd -										0.0097	1	0.27																															.021
hist_user_behavior_n_seekback											0.27	1	0.026																														0092
hist_user_behavior_is_shuffle												0.026	1	0.0079																													.071
hour_of_day -													0.0079	1	0.013																					1.024							.0035
premium -														0.013	1	0.082																											-
															10/10/20		1000000																				No. of Contract of			1000000		0.015 -0	.053
release_year -																		1000																		_	10000	0.28				-	1.47
us_popularity_estimate -																	-		STATE OF THE PARTY											-	2.00											0.13 0	
acousticness -																		and the same of		Contract of the Contract of th										_						-			_			0.19	0.26
beat_strength -																			100000				100000	1000						10000	-	10000			200				1000			0.16	1.26
danceability -																							100000000						-	-	and Administration of											-0.12	
dyn_range_mean -																			1000				E0000016	the state of					STATE OF TAXABLE PARTY.	ROCK I					0.00	-	NAME OF TAXABLE PARTY.		200000			-0.12	
																								September 1				_		- NAVA-							-	- Control of the Cont	-			0.11	
flatness																						_			5 3 7			-							-		-	0.35				0.088	0.073
instrumentalness -																	-0.066			0.15	-0.17	0.2	0.14	0.11	0.054	1	-0.0068			0.047					0.11	0.33	0.41	0.34	0.033				0.14
key -																										0.0068	1								0.085	.011							.0081
liveness -																											-0.022	1															0.017
loudness -																								0.69				0.037	1	0.27						0.22	0.56						0.15
mechanism -																			-0.37			0.62							0.27	1	-0.86					0.25							0.74
organism -																		0.006	0.77	0.4	0.32	0.51	0.28						0.39	-0.86	1	0.093			0.13	0.31	0.41		0.045			0.13	0.28
speechiness -																				-													NAME OF TAXABLE PARTY.									0.31	10000
tempo -																																No. of Concession,											
time_signature -																								_										_	-				WANTED BY	-		-0.16 0	NAME OF TAXABLE PARTY.
																								CONTRACT.												NAME OF STREET		700000	CONTRACT OF THE PARTY OF	Market 1	0.054	0.28	66.0000
acoustic_vector_0																										_					-					-	-0.47	-0.54		0.58	0.41	0.32	-0.58
acoustic_vector_1 -																	CANADA S				1000000			A STATE OF THE PARTY OF T						W. C.							1	0.37	-0.24		0.33	0.11	1.42
acoustic_vector_2 - acoustic vector 3 -																			1000000									0.14									0.37	0.43	0.41	0.36	0.70		0.10
acoustic vector 3 -																				100000	1/800001	200	100000000000000000000000000000000000000								_	000000000000000000000000000000000000000				The state of the s		-0.36	0.042	1	0.022	0.37	0.61
acoustic_vector_4 -																																		_	_		-		-	-0.022	1	0.045 -0	0066
acoustic vector 6																																				0.000	COLUMN TO SERVICE STATE OF THE PERSON SERVICE STATE OF THE	nessinger:			0.045	1	0.051
acoustic_vector_7 -																	_														-						0.42		NOT THE	-0.61	.0066 -	0.051	1
	5	-6	- 5	2	re,	· p	6	ay -	- ke	ay -	9	Ą.	ė	ay -	É	5	- 10	ė.	i	É	ė	- Au	- ue	- AG	ż	18	, i	ė	ė	Ė	E.	2	ž.	- Ju	ś	0	ń	7	m	w.	ń	6	1
	positi	leng	skip	skip	dops	ddiy	t swit	are pl	d al	are pl	seekth	ekba	Shuff	p o	remin	durati	95	Stime	sticne	treng	incine	inge	e a	ě	flatme	taine	.2	byene	andne	chanis	ridenis	schine	E	gnatt	valen	vector	vector	rector	vector	vector	vector	vector	vector
	Sion	ession				5	untext	2	Z	Def.	5	5	iarja	hour	ď.	Ĩ.	relea	rity_e	acon	Deat.	ğ	S S	Gue			umen		(7%)	24	ŠE	8	8		E S		stic_3	abc.,	apc.)	200	SDC )	Stic v	atic_v	Spir.
	8	28					8	35	3	31	3	50	20					9		au .			5			43								40		8	70	20	3	20	8	8	8

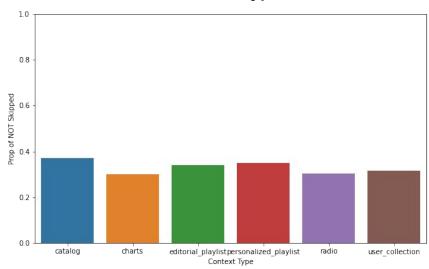
- 0.75

- -0.50

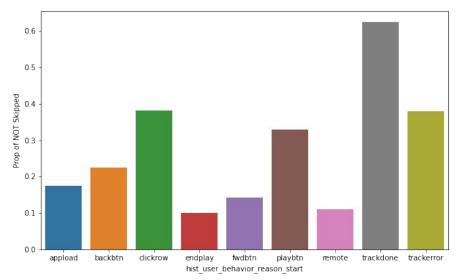
-- 0.7

#### **EDA - Examine Multi-class Categorical Variables**

## Proportion of Non-Skipped Songs by Context Types



# Proportion of Non-Skipped Songs by Play Reason



Flat pattern indicates weak predictor

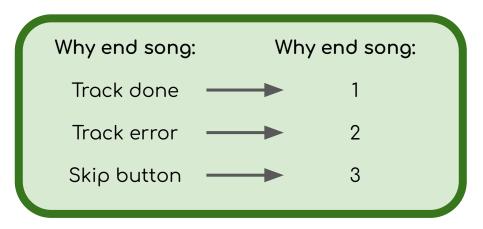
**Sharp Contrast** indicates **strong influencer** 

#### Complex variables

- skip\_1, skip\_2, skip\_3, not\_skipped
  - Target variables that all depend on each other
- Categorical variables, e.g. "why the user started the song", "why the user ended the song"
  - o Possible values: "track done", "track error", "skip button", etc.

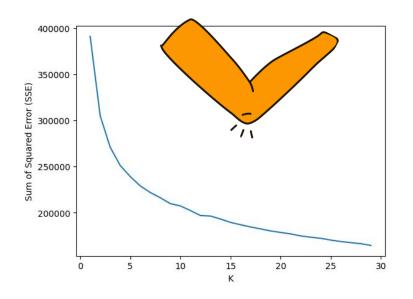
#### Approaches for categorical variables:

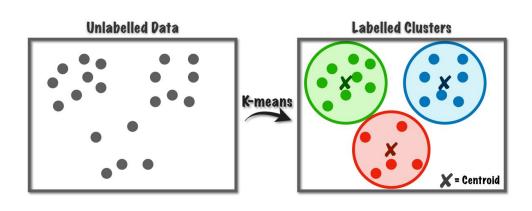
- a. Label encoding (hierarchical)
- b. One-hot encoding (dummy encoding)
- c. Exclude entirely



#### **Unsupervised learning: K-means Clustering**

- 2 clusters: accuracy = 42%
- No sharp elbow
- 30 clusters: clustered based on instrumental features

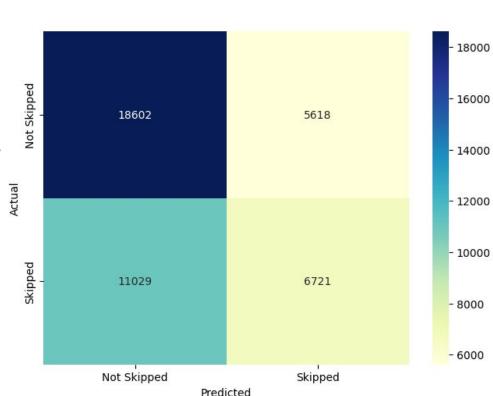




#### Logistic Regression: Two Approaches

- 1st model: Logistic regression using only numeric variables (no categorical variables)
- Confusion matrix displays results of all trials as a heatmap
  - Ex) 18602 runs correctly predicted non skipped song as non skipped

Accuracy score: 0.6034



Confusion matrix

#### Logistic Regression: Two Approaches

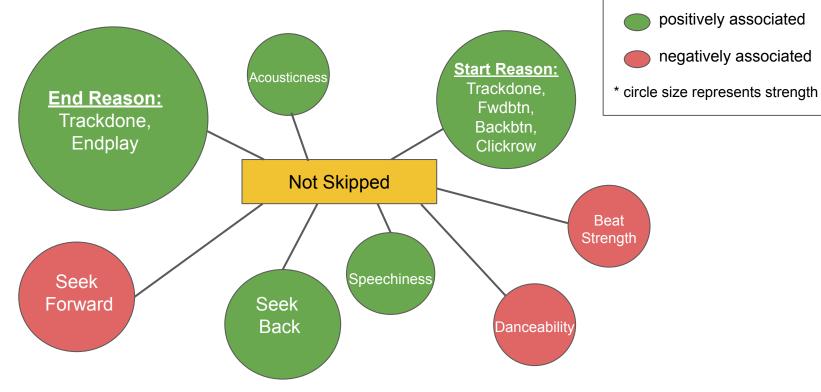
- Change in feature selection: dummy encoding the "strong influencers" and remove the "weak predictors"
- Prediction Accuracy on test set:

98.8%





#### Top Important Features based on Coefficients

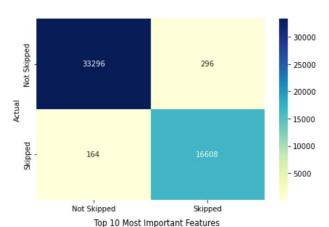


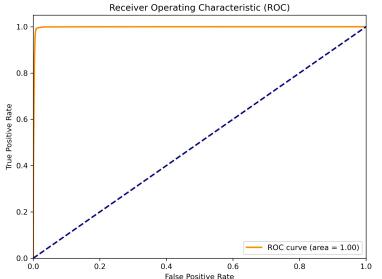
#### Confusion matrix

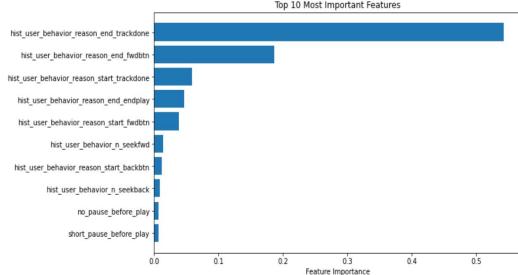
#### **Random forest - Better Predictions**

- Applied random forest classification with 100 estimators
- Model has strong ability to classify not-skipped vs. skipped — ROC curve

99.1% prediction Accuracy on test set







#### **Conclusion**

- 3 Machine Learning Techniques:
  - Logistic Regression w/ & w/o categorical variables
  - Random Forest
  - Unsupervised Learning
- Logistic Regression Model with selected categorical variables had the highest accuracy: 0.9908
  - Including lowly correlated variables created a stronger model than only including strongly correlated variables
- Better understanding of the data collected by Spotify can help build stronger models
  - Ex) Applying non-linear relationships between variables



#### Links and sources

Headphones image: <a href="https://www.publicdomainpictures.net/en/view-image.php?image=212667&picture=black-headphones">https://www.publicdomainpictures.net/en/view-image.php?image=212667&picture=black-headphones</a>

K-means diagram: <a href="https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c">https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c</a>