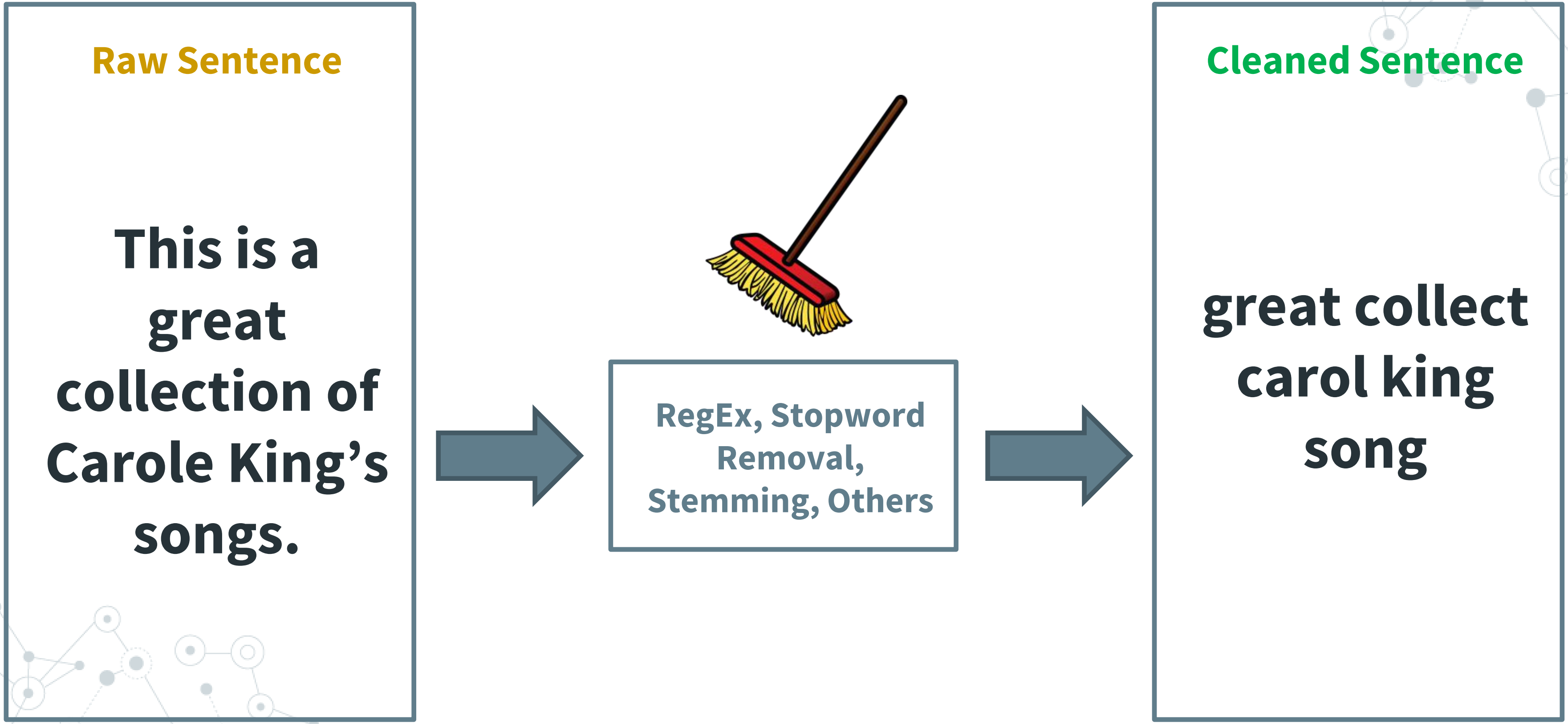
amazon music

Rating Prediction

By Fendi Halim Tjoa & MIE1624 Group 7

Data Cleaning – Text Data



Raw Sentence

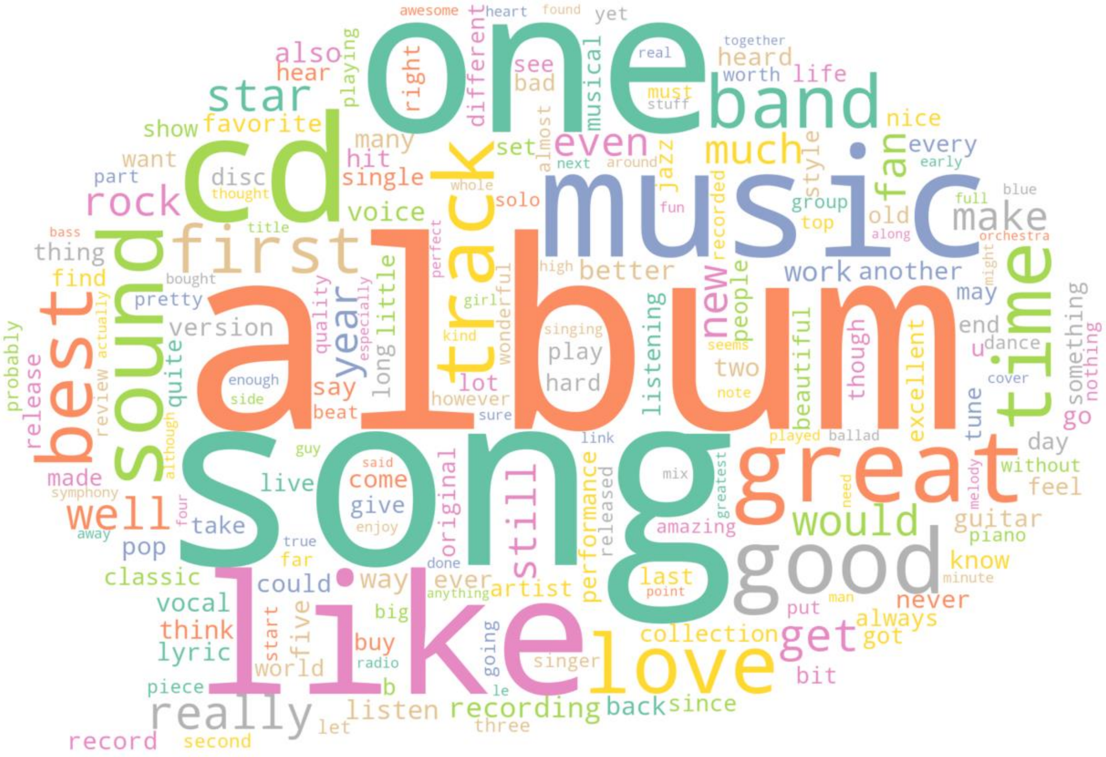
This is a great collection of Carole King’s songs.

RegEx, Stopword Removal, Stemming, Others

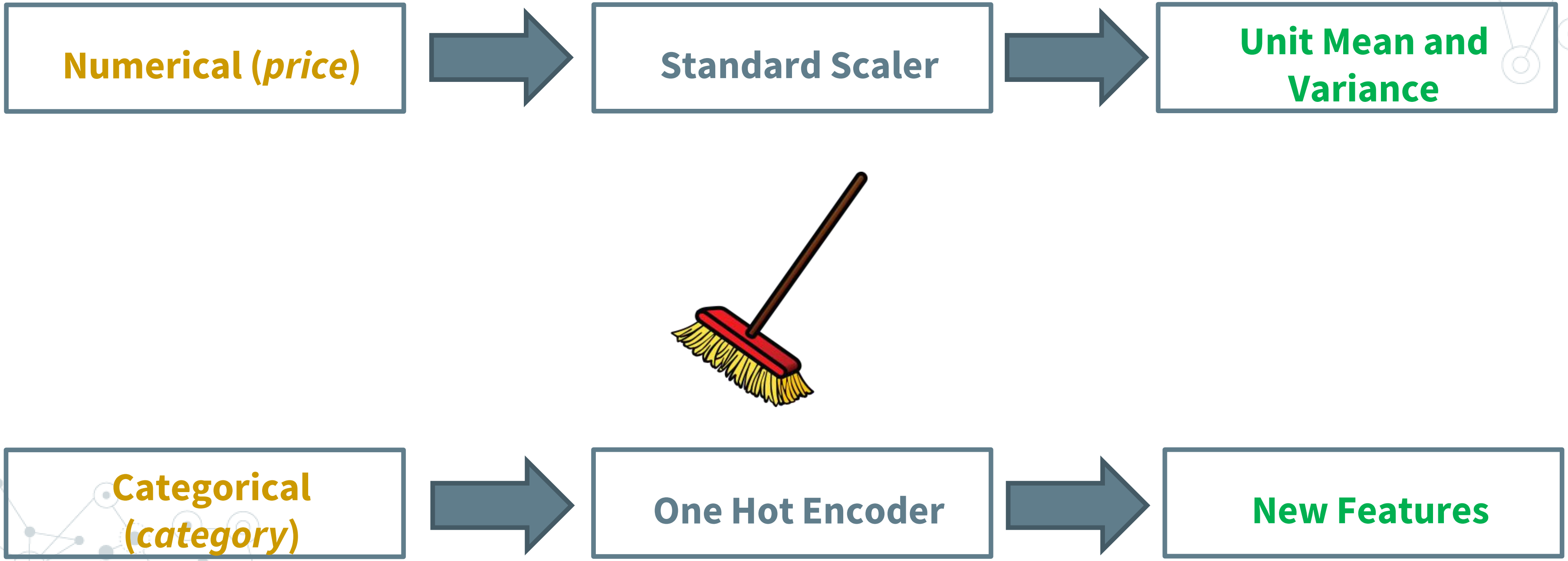
Cleaned Sentence

great collect carol king song

Cleaned Text - Word Cloud Visualization



Data Cleaning – Numerical and Categorical Data



Numerical (*price*)

Standard Scaler

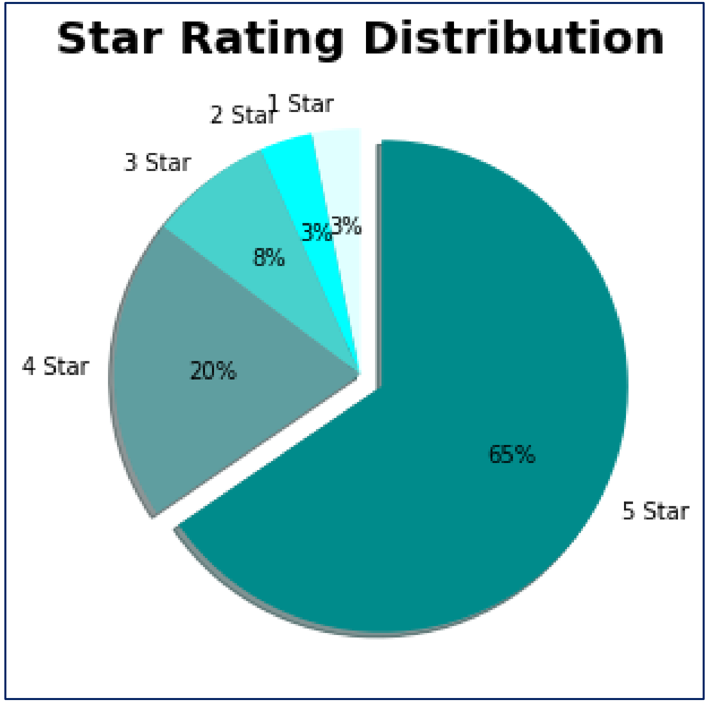
Unit Mean and Variance

Categorical (*category*)

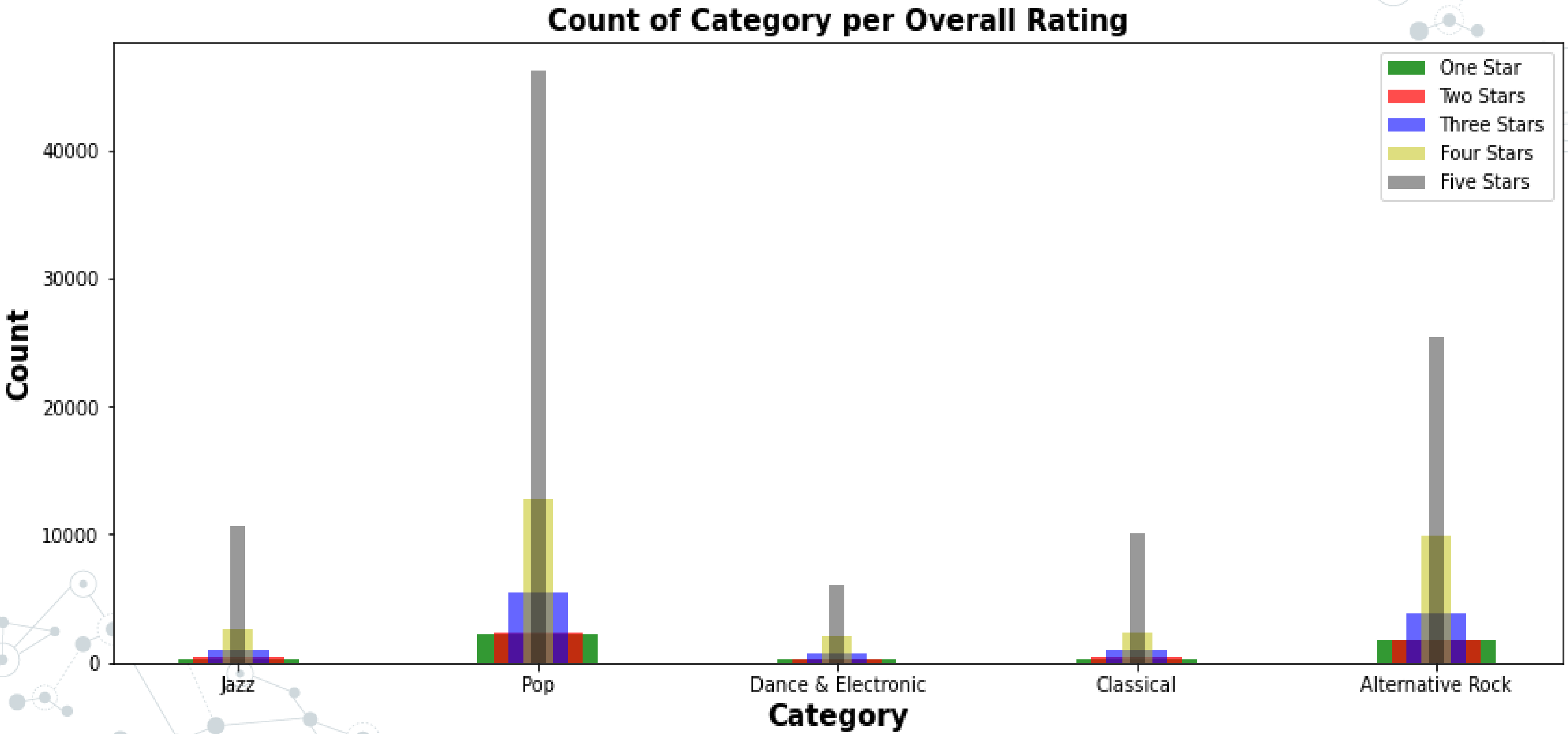
One Hot Encoder

New Features

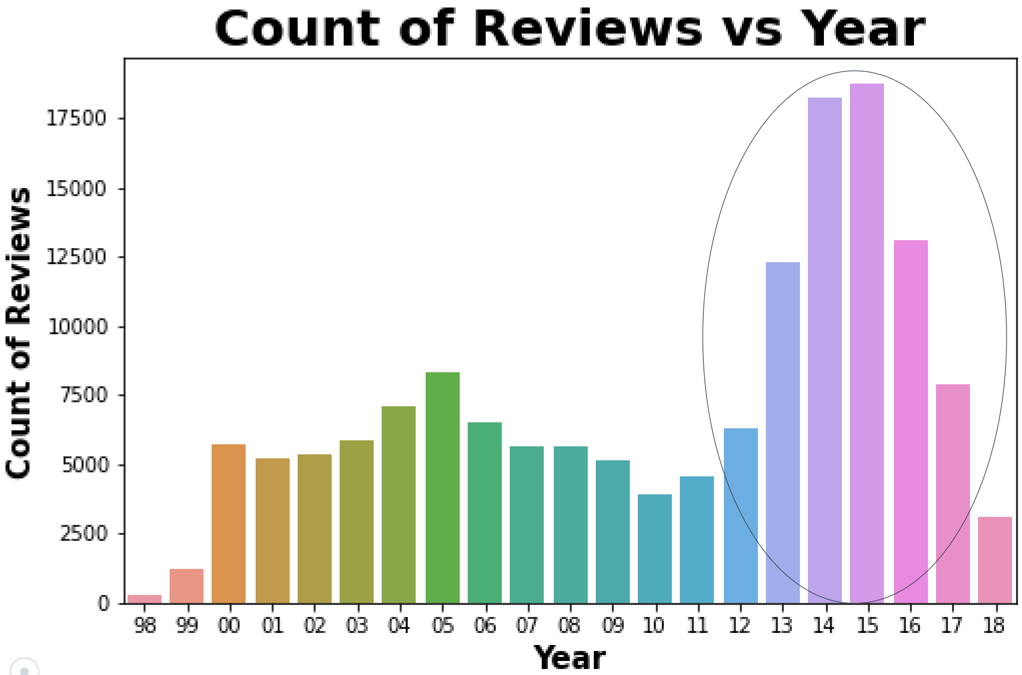
Exploratory Data Analysis (EDA) – Target Distribution



EDA – Target Distribution vs Category



EDA – Review Counts vs Year

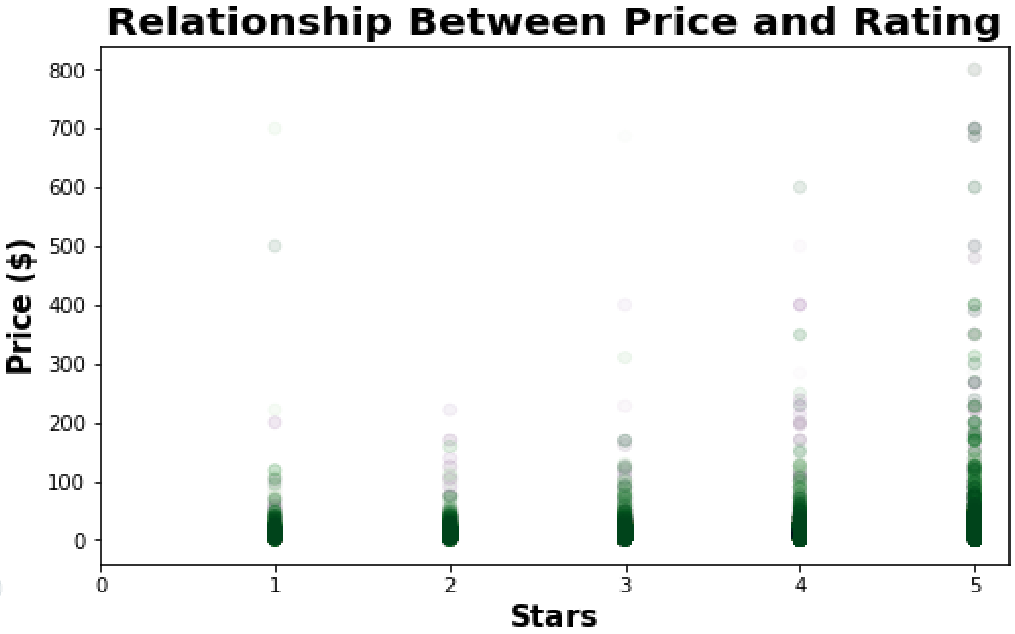


New Era of Smart Phone

Smartphone booming

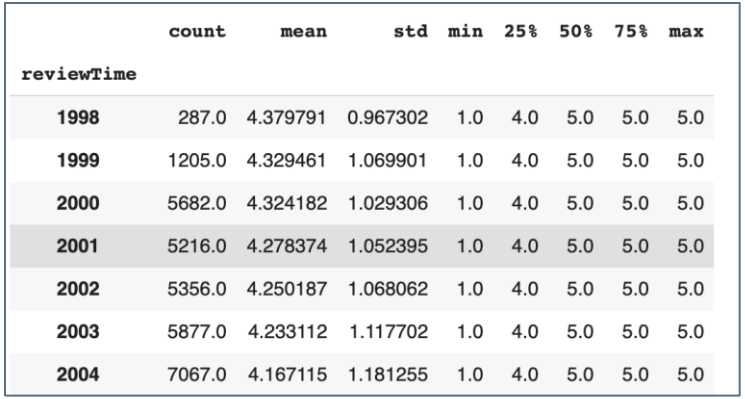
Feature Importance

- Explanatory feature: Price

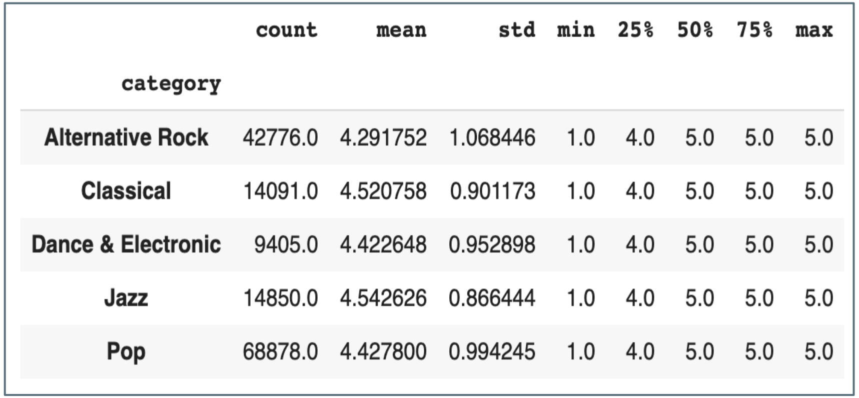


Feature Importance

- Less significant features: UnixTime, category

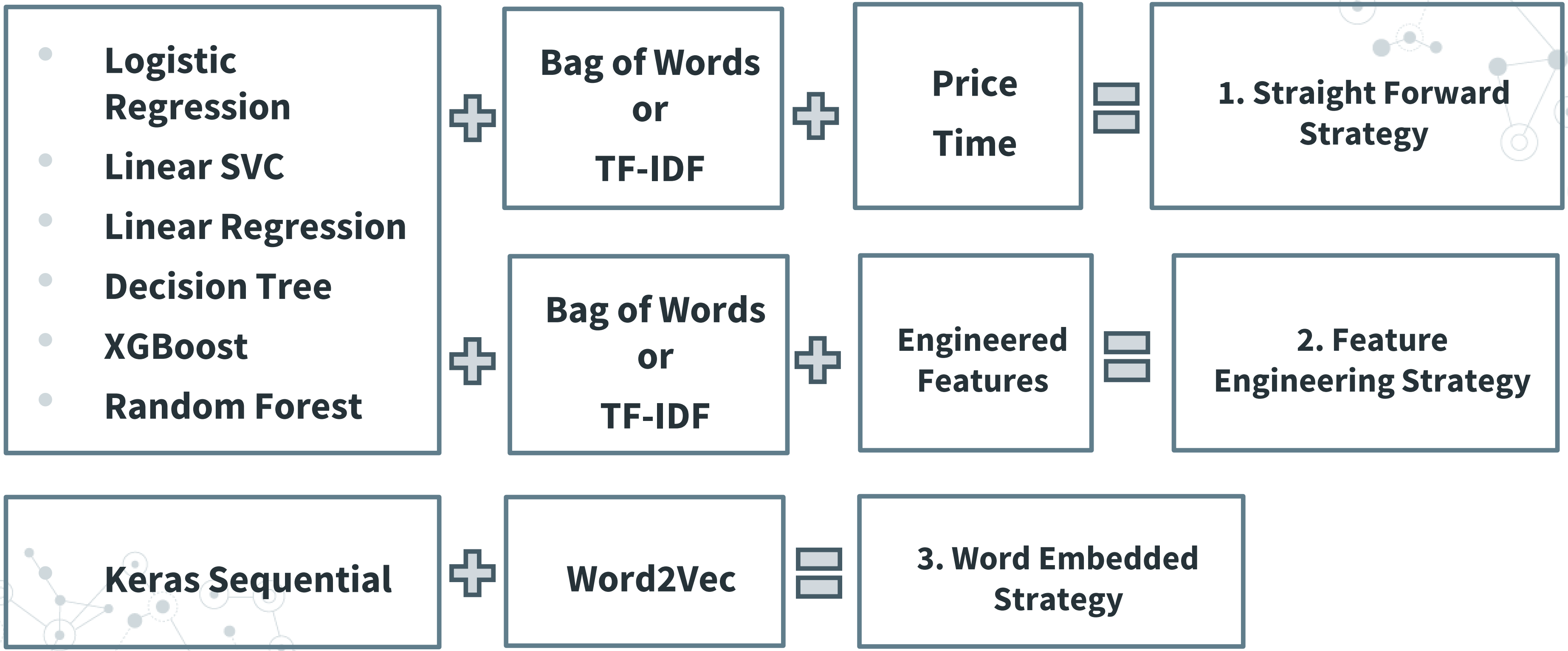


	count	mean	std	min	25%	50%	75%	max
reviewTime								
1998	287.0	4.379791	0.967302	1.0	4.0	5.0	5.0	5.0
1999	1205.0	4.329461	1.069901	1.0	4.0	5.0	5.0	5.0
2000	5682.0	4.324182	1.029306	1.0	4.0	5.0	5.0	5.0
2001	5216.0	4.278374	1.052395	1.0	4.0	5.0	5.0	5.0
2002	5356.0	4.250187	1.068062	1.0	4.0	5.0	5.0	5.0
2003	5877.0	4.233112	1.117702	1.0	4.0	5.0	5.0	5.0
2004	7067.0	4.167115	1.181255	1.0	4.0	5.0	5.0	5.0



	count	mean	std	min	25%	50%	75%	max
category								
Alternative Rock	42776.0	4.291752	1.068446	1.0	4.0	5.0	5.0	5.0
Classical	14091.0	4.520758	0.901173	1.0	4.0	5.0	5.0	5.0
Dance & Electronic	9405.0	4.422648	0.952898	1.0	4.0	5.0	5.0	5.0
Jazz	14850.0	4.542626	0.866444	1.0	4.0	5.0	5.0	5.0
Pop	68878.0	4.427800	0.994245	1.0	4.0	5.0	5.0	5.0

Model Implementation - Strategies



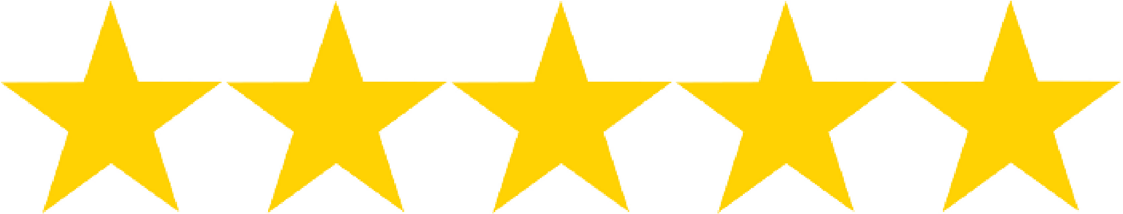
- Logistic Regression
- Linear SVC
- Linear Regression
- Decision Tree
- XGBoost
- Random Forest

+ Bag of Words or TF-IDF + Price Time = 1. Straight Forward Strategy

+ Bag of Words or TF-IDF + Engineered Features = 2. Feature Engineering Strategy

Keras Sequential + Word2Vec = 3. Word Embedded Strategy

Classification vs Regression

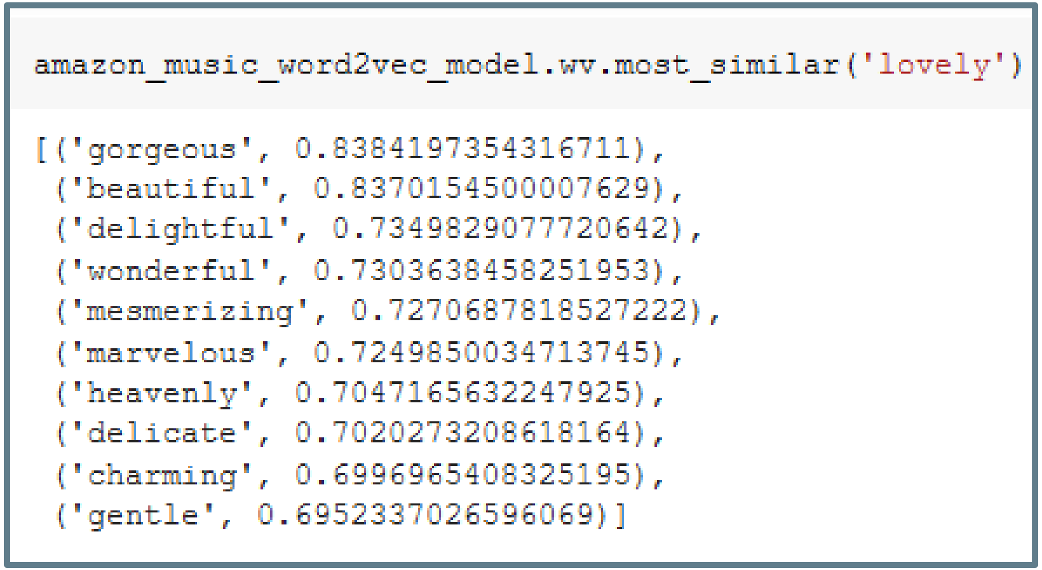


Discrete

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)}_{\text{The square of the difference between actual and predicted}}^{2}$$

- Utilizing Linear Regression to minimize MSE
- Continuous Output

Word Embedding

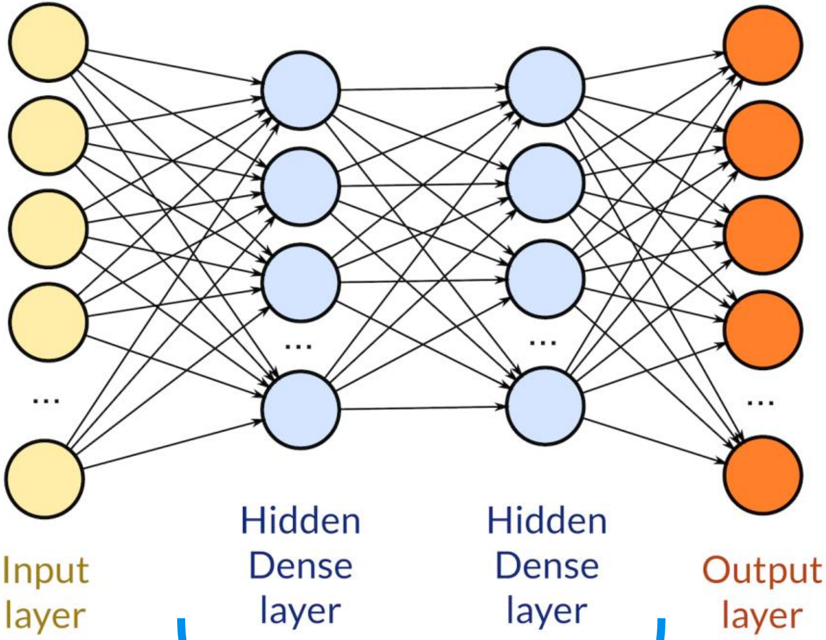


```
amazon_music_word2vec_model.wv.most_similar('lovely')
```

```
[('gorgeous', 0.8384197354316711),
 ('beautiful', 0.8370154500007629),
 ('delightful', 0.7349829077720642),
 ('wonderful', 0.7303638458251953),
 ('mesmerizing', 0.7270687818527222),
 ('marvelous', 0.7249850034713745),
 ('heavenly', 0.7047165632247925),
 ('delicate', 0.7020273208618164),
 ('charming', 0.6996965408325195),
 ('gentle', 0.6952337026596069)]
```

Word2Vec

Keras Sequential



Too many unexplored layers

Incorrect settings + too little epochs

Underfitting

Training = High MSE

Validation = High MSE

Incorrect settings + too many epochs

Overfitting

Training = Low MSE

Validation = High MSE

Results and Discussion

	Model	Training MSE (Straight Forward)	Validation MSE (Straight Forward)	Training MSE (Feature Engineered)	Validation MSE (Feature Engineered)
1	Logistic Regression	0.6184	0.666	0.665	0.721
2	Linear SVC	0.5286	0.69	0.883	1.038
3	Linear Regression	0.4571	0.516	0.579	0.641
4	Decision Tree	0	1.174	0	1.185
5	XGBoost	0.949	0.999	0.883	0.917
6	Random Forest	0.0004	1.144	0.0003	1.178

	Model	Training MSE	Validation MSE
1	Keras Sequential	0.347475	1.1447

Thank You!

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash

References:

- https://www.dataquest.io/blog/understanding-regression-error-metrics/
- https://medium.com/@carmensample/thank-you-for-the-one-star-review-907f93d08a0b
- https://www.businessinsider.com.au/idc-the-smartphone-boom-is-over-2016-9