Amazon Automotive Products Recommender System

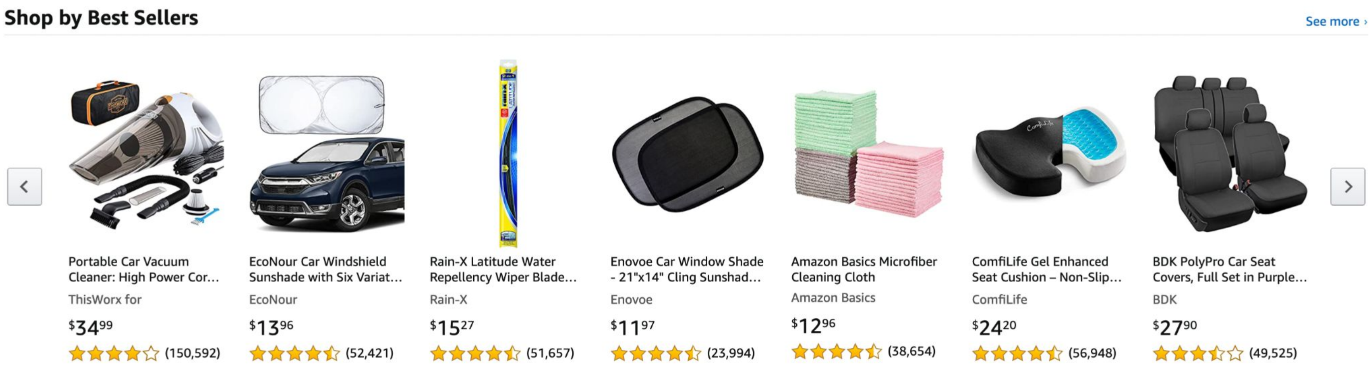
By: Haasitha Pidaparthi

INTRODUCTION

- Most online e-commerce sites use some type of recommendation system
    - Amazon uses item-based collaborative filtering
- Automotive products generally have the most reviews on Amazon
- Gain insight on Amazon’s use of review data for their recommendation system

CHALLENGE

- Build a Recommender System for Amazon Automotive Products
- Compare and find the best recommender system algorithm



Shop by Best Sellers

See more ›

Portable Car Vacuum Cleaner: High Power Cor…
ThisWorx for
$34.99
(150,592)

EcoNour Car Windshield Sunshade with Six Variat…
EcoNour
$13.96
(52,421)

Rain-X Latitude Water Repellency Wiper Blade…
Rain-X
$15.27
(51,657)

Enovoe Car Window Shade - 21"x14" Cling Sunshad…
Enovoe
$11.97
(23,994)

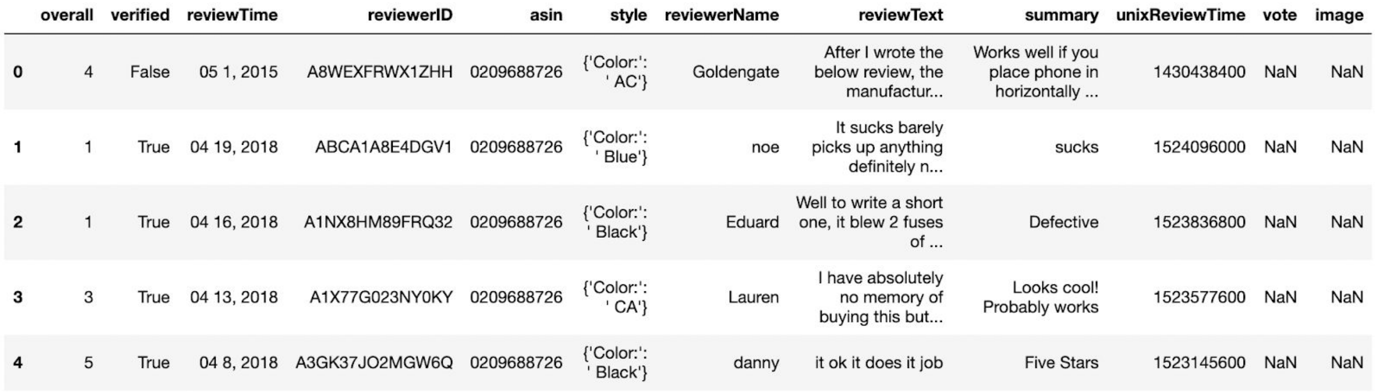
Amazon Basics Microfiber Cleaning Cloth
Amazon Basics
$12.96
(38,654)

ComfiLife Gel Enhanced Seat Cushion – Non-Slip…
ComfiLife
$24.20
(56,948)

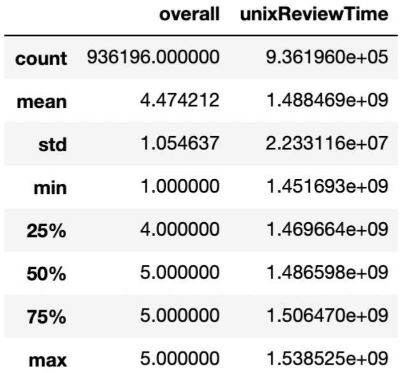
BDK PolyPro Car Seat Covers, Full Set in Purple…
BDK
$27.90
(49,525)

DATASET

“Small” subset of 5-core reviews (~1.7 million rows), 193 MB



	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	image
0	4	False	05 1, 2015	A8WEXFRWX1ZHH	0209688726	{'Color:': ' AC'}	Goldengate	After I wrote the below review, the manufactur...	Works well if you place phone in horizontally ...	1430438400	NaN	NaN
1	1	True	04 19, 2018	ABCA1A8E4DGV1	0209688726	{'Color:': ' Blue'}	noe	It sucks barely picks up anything definitely n...	sucks	1524096000	NaN	NaN
2	1	True	04 16, 2018	A1NX8HM89FRQ32	0209688726	{'Color:': ' Black'}	Eduard	Well to write a short one, it blew 2 fuses of ...	Defective	1523836800	NaN	NaN
3	3	True	04 13, 2018	A1X77G023NY0KY	0209688726	{'Color:': ' CA'}	Lauren	I have absolutely no memory of buying this but...	Looks cool! Probably works	1523577600	NaN	NaN
4	5	True	04 8, 2018	A3GK37JO2MGW6Q	0209688726	{'Color:': ' Black'}	danny	it ok it does it job	Five Stars	1523145600	NaN	NaN



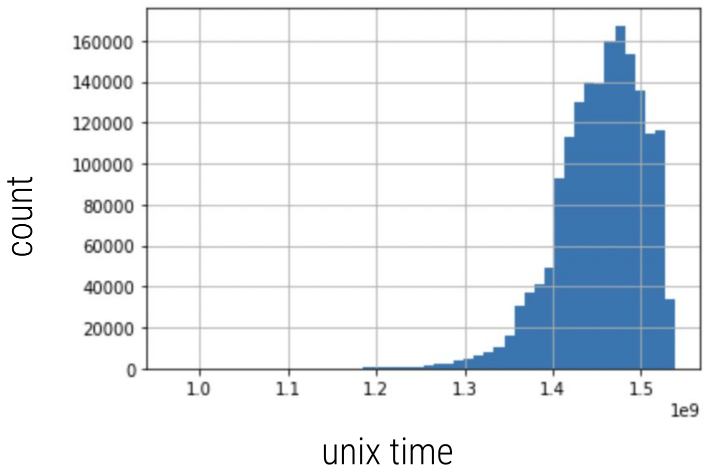
	overall	unixReviewTime
count	936196.000000	9.361960e+05
mean	4.474212	1.488469e+09
std	1.054637	2.233116e+07
min	1.000000	1.451693e+09
25%	4.000000	1.469664e+09
50%	5.000000	1.486598e+09
75%	5.000000	1.506470e+09
max	5.000000	1.538525e+09

Dataset Link: https://nijianmo.github.io/amazon/index.html#subsets

PREPROCESSING STEPS

1. Sort value by UnixReviewTime
2. Dropped reviews older than 2016 (~740,500)
3. Analyze verified column
4. Dropped reviewerName, reviewText, summary, image
5. Considered vote and style columns
   a. Missing a lot of data

Unix Review Time

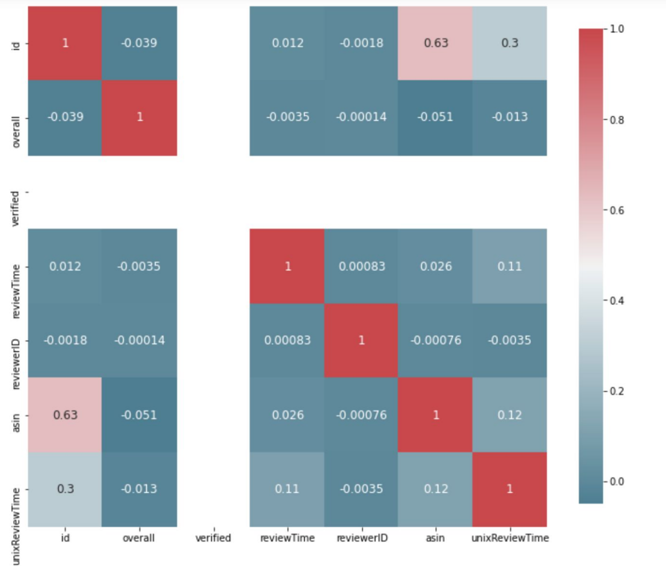


1

SCIKIT-LEARN

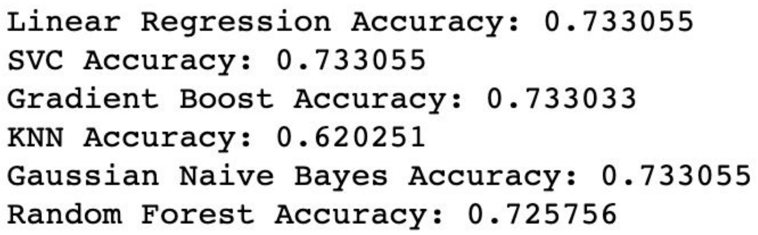
SCIKIT-LEARN

- Linear Regression, SVM, Gradient Boost, KNN, Gaussian Naive Bayes, and Random Forest algorithms
- Correlation map to check target relevancy
- Label Encoding on non-numeric columns

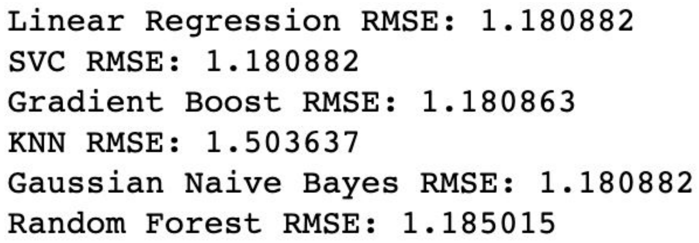


EVALUATION

RMSE and Accuracy metrics:



```
Linear Regression Accuracy: 0.733055
SVC Accuracy: 0.733055
Gradient Boost Accuracy: 0.733033
KNN Accuracy: 0.620251
Gaussian Naive Bayes Accuracy: 0.733055
Random Forest Accuracy: 0.725756
```



```
Linear Regression RMSE: 1.180882
SVC RMSE: 1.180882
Gradient Boost RMSE: 1.180863
KNN RMSE: 1.503637
Gaussian Naive Bayes RMSE: 1.180882
Random Forest RMSE: 1.185015
```

2

SURPRISE

SURPRISE

- Algorithms used:
  - KNNBasic (KNN based; User-based Collaborative filtering)
  - SVD (matrix factorization)
  - Co-Clustering (cluster based)
  - Slope One (Item-based Collaborative filtering)
- Custom devised time-based KNN: Utilizes Surprise get_neighbors method and UnixReviewTime column.

EVALUATION

k	RMSE (Time Based)	RMSE (Surprise)
3	1.0265	0.9942
4	1.0035	0.9916
5	0.9880	0.991
6	0.978	0.9663
7	0.973	0.9662
8	0.9674	0.9660
9	0.9665	0.9659

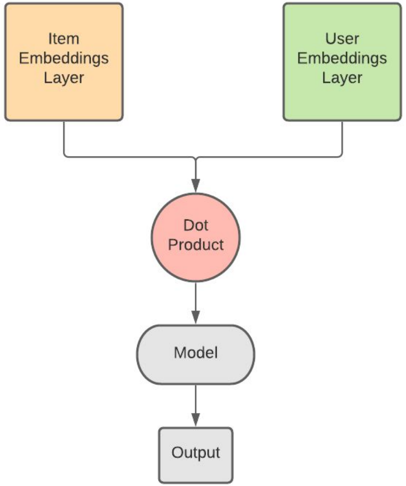
Algorithm	RMSE
SVD	0.9916
Co-Clustering	1.0849
Slope One	1.0323
KNN	0.9736

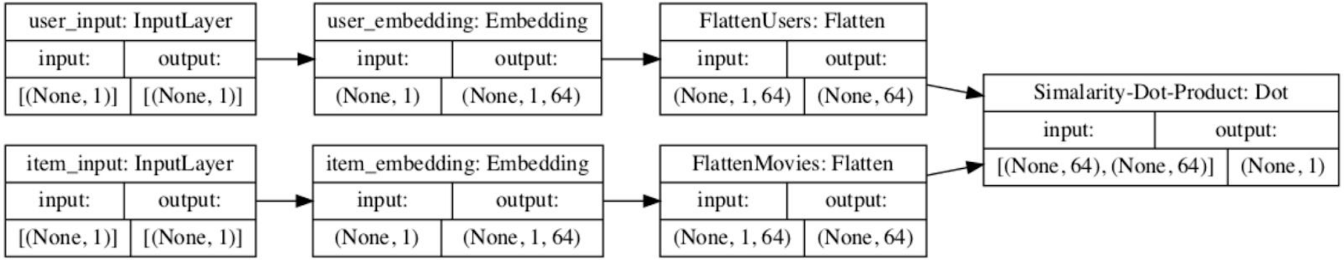
3

TENSORFLOW KERAS

CONVOLUTIONAL NEURAL NETWORK (CNN)

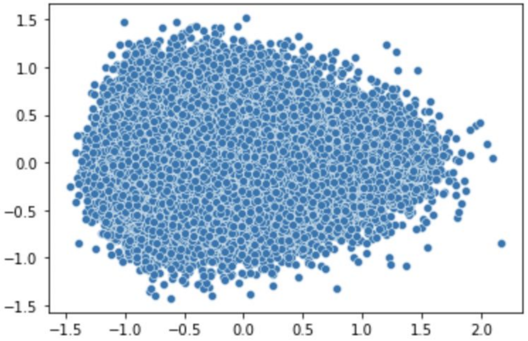
- Input: Input for both items and users
- Embedding Layers: Embeddings for items and users
- Dot: combines embeddings using a dot product

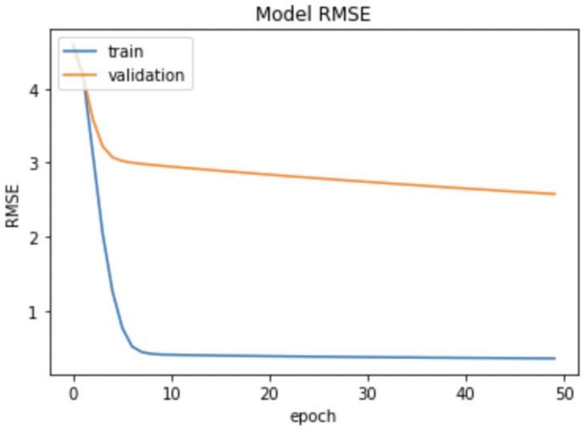


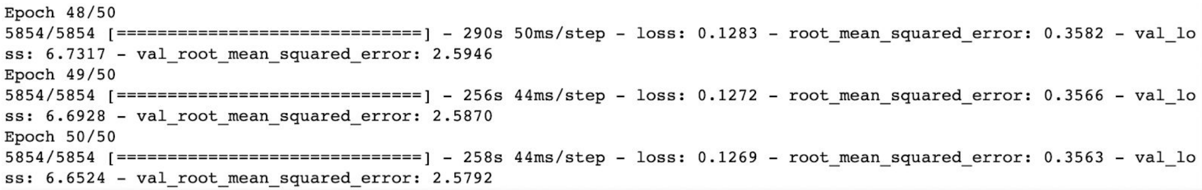


EVALUATION

Embeddings Visualization







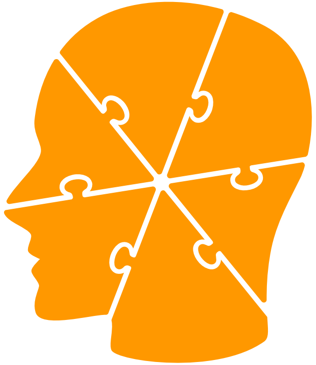
```
Epoch 48/50
5854/5854 [==============================] - 290s 50ms/step - loss: 0.1283 - root_mean_squared_error: 0.3582 - val_lo
ss: 6.7317 - val_root_mean_squared_error: 2.5946
Epoch 49/50
5854/5854 [==============================] - 256s 44ms/step - loss: 0.1272 - root_mean_squared_error: 0.3566 - val_lo
ss: 6.6928 - val_root_mean_squared_error: 2.5870
Epoch 50/50
5854/5854 [==============================] - 258s 44ms/step - loss: 0.1269 - root_mean_squared_error: 0.3563 - val_lo
ss: 6.6524 - val_root_mean_squared_error: 2.5792
```

CONCLUSION

- Surprise KNN had the best RMSE of 0.9658
  - But only ran on ~50,000 elements of the dataset
- SVD had the best overall results with RMSE of 0.9823
  - HyperParameters values
    - epochs: 30, lr_all = 0.005, reg_all = 0.05, factors = 20
- Surprise KNN and SciKit-Learn KNN had contradictory results
  - had some similarities in run time and the strain on local resources

FURTHER WORK

- Natural Language Processing (NLP) on user reviews
- Merge data from different sources
- Try overall recommendation on all categories



Thank you!

Any Questions?