Subreddit Classification & Business Implications

Amanda Jo Russell
GA • DSI-7 • Project 3
04.05.2019

Data Science Problem: *Which classification model most accurately predicts subreddit origin?*

r/ Uber

- Launched: March 2009
- Available in 65 countries / 600 cities
- Rides per day: 15 million
- Valuation: $72 billion
- Subsidiaries
  - Uber Eats
  - Jump Bikes

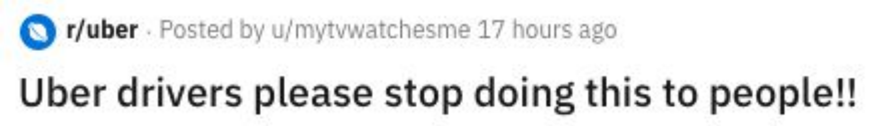
r/ lyft

- Launched: June 2012
- Available in 2 countries / 300 cities
- Rides per day: 1 million
- Valuation: $15 billion
- Additional Ventures
  - Walt Disney World Resort: “Minnie Van“
  - Scooter & Bike-sharing

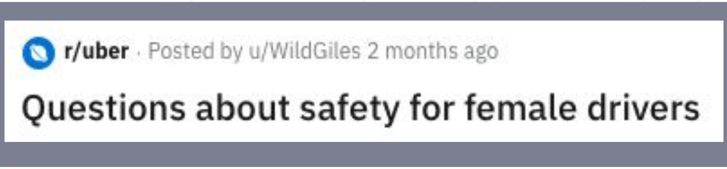
Source: http://www.businessofapps.com/data/uber-statistics

Data Collection: Examples

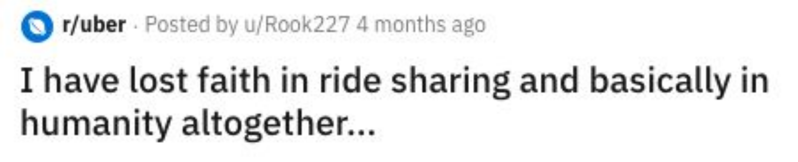
1401 Total Posts (701 Lyft / 700 Uber)


r/uber · Posted by u/mytvwatchesme 17 hours ago

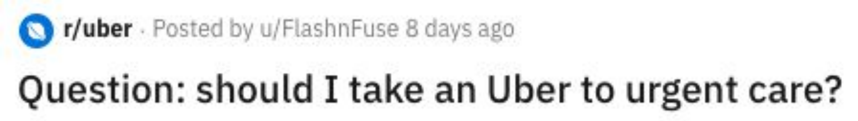
Uber drivers please stop doing this to people!!


r/uber · Posted by u/WildGiles 2 months ago

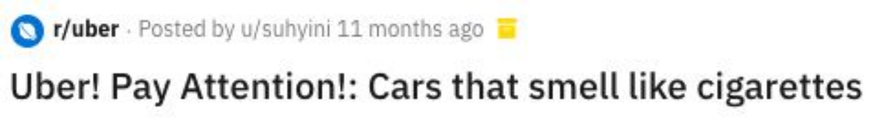
Questions about safety for female drivers


r/uber · Posted by u/Rook227 4 months ago

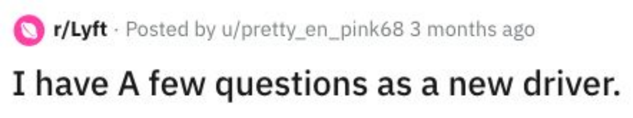
I have lost faith in ride sharing and basically in humanity altogether...


r/uber · Posted by u/FlashnFuse 8 days ago

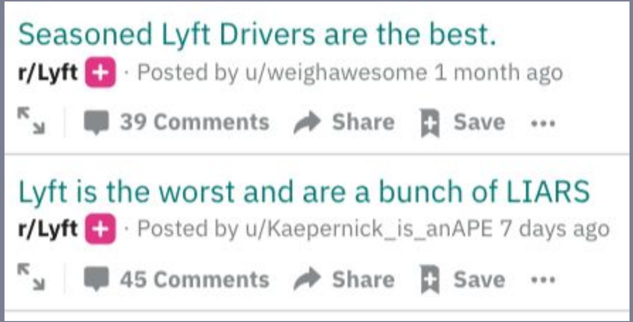
Question: should I take an Uber to urgent care?


r/uber · Posted by u/suhyini 11 months ago

Uber! Pay Attention!: Cars that smell like cigarettes


r/Lyft · Posted by u/pretty_en_pink68 3 months ago

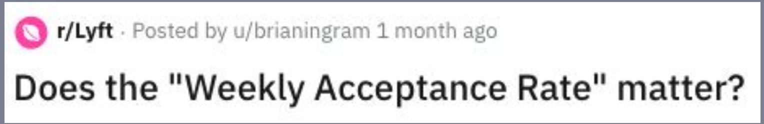
I have A few questions as a new driver.


Seasoned Lyft Drivers are the best.

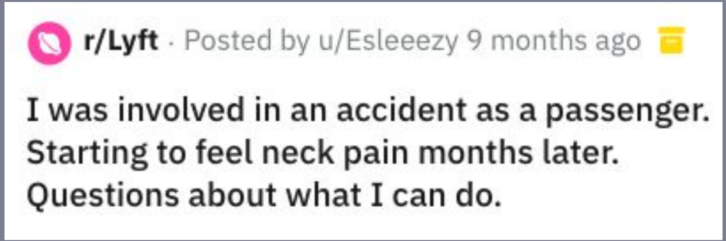
r/Lyft · Posted by u/weighawesome 1 month ago

Lyft is the worst and are a bunch of LIARS

r/Lyft · Posted by u/Kaepernick_is_anAPE 7 days ago

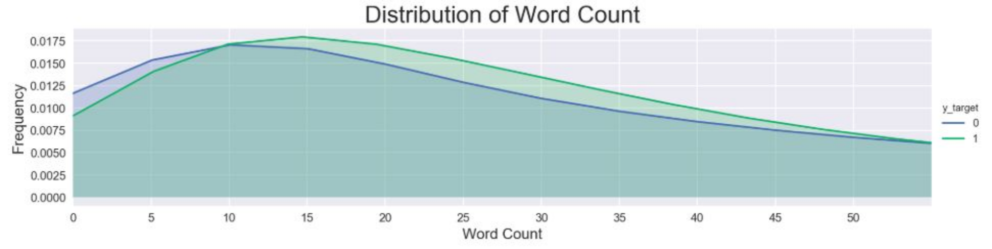

r/Lyft · Posted by u/brianingram 1 month ago

Does the "Weekly Acceptance Rate" matter?


r/Lyft · Posted by u/Esleeezy 9 months ago

I was involved in an accident as a passenger. Starting to feel neck pain months later. Questions about what I can do.

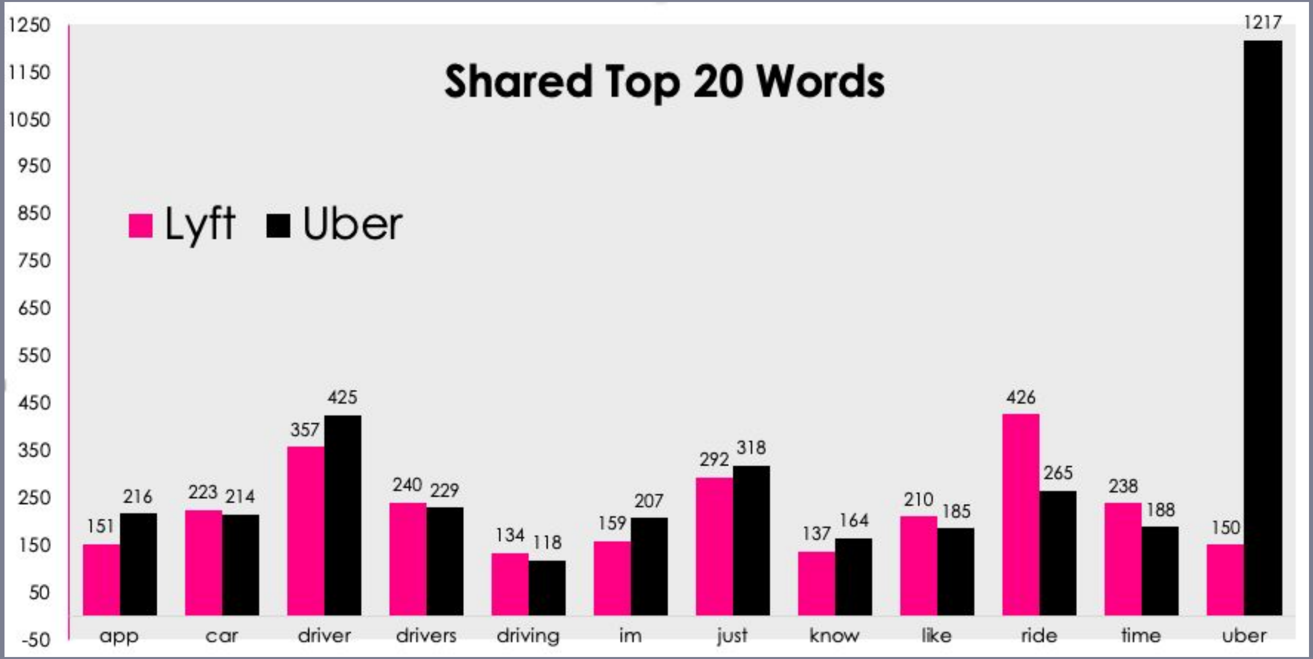
Initial Observations



Unique Top 20 Words:

<u>Lyft</u>
#1: lyft
#9: rides
#10: passenger
#11: drive
→ ***#15: uber***

<u>Uber</u>
#9: account
#12: got
#14: don't
#15: card
→ *NOT lyft*



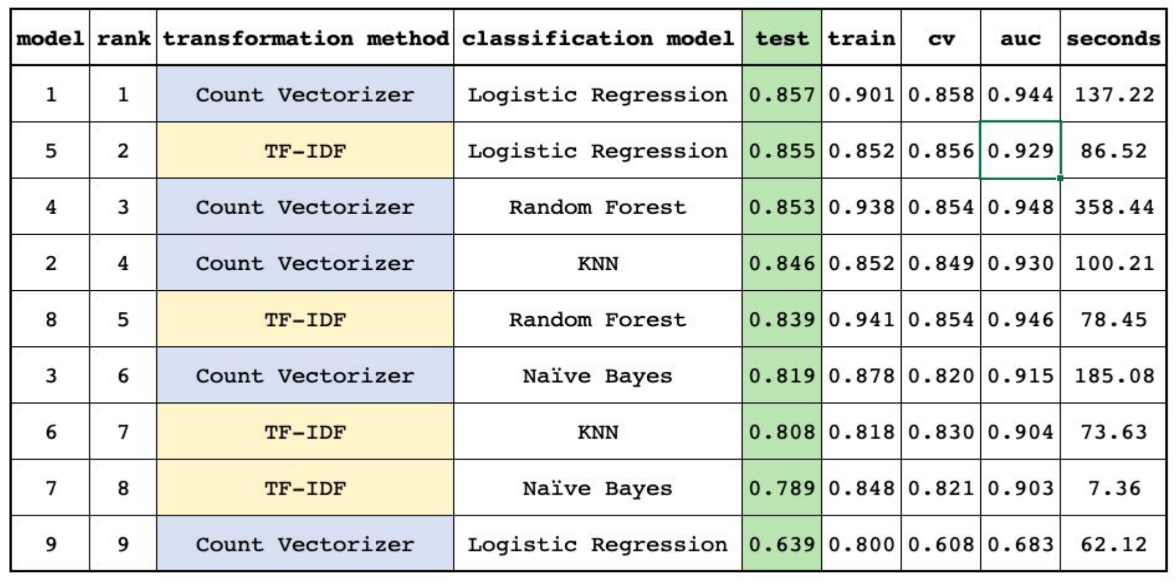
Models & Performance

Transformation Methods

- Count Vectorization*
- TF-IDF

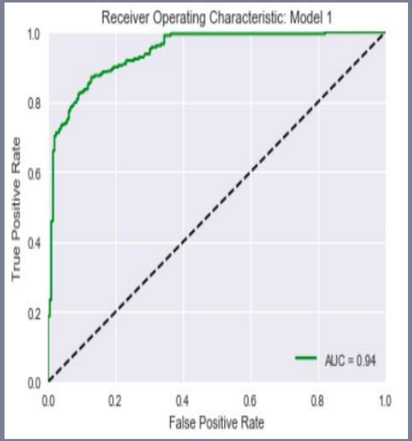
Classification Models

- Logistic Regression*
- KNN
- Naive Bayes
- Random Forest

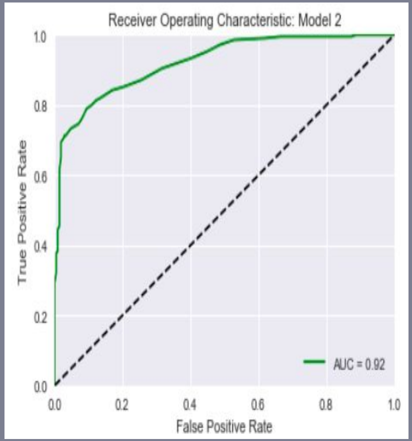


model	rank	transformation method	classification model	test	train	cv	auc	seconds
1	1	Count Vectorizer	Logistic Regression	0.857	0.901	0.858	0.944	137.22
5	2	TF-IDF	Logistic Regression	0.855	0.852	0.856	0.929	86.52
4	3	Count Vectorizer	Random Forest	0.853	0.938	0.854	0.948	358.44
2	4	Count Vectorizer	KNN	0.846	0.852	0.849	0.930	100.21
8	5	TF-IDF	Random Forest	0.839	0.941	0.854	0.946	78.45
3	6	Count Vectorizer	Naïve Bayes	0.819	0.878	0.820	0.915	185.08
6	7	TF-IDF	KNN	0.808	0.818	0.830	0.904	73.63
7	8	TF-IDF	Naïve Bayes	0.789	0.848	0.821	0.903	7.36
9	9	Count Vectorizer	Logistic Regression	0.639	0.800	0.608	0.683	62.12

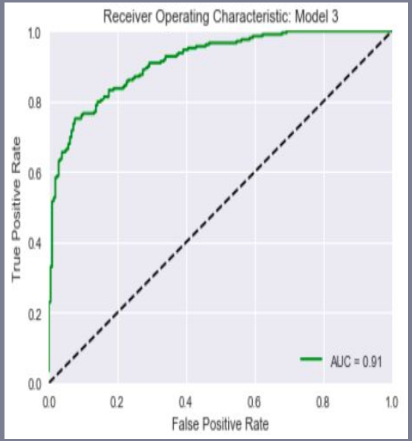
Further Exploration: adjust hyperparameters of existing models, additional transformation methods and/or classification models, add/remove stop words



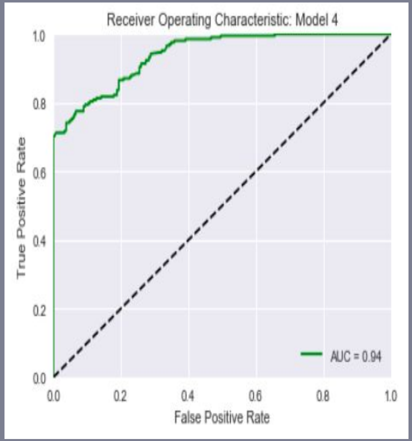
Receiver Operating Characteristic: Model 1



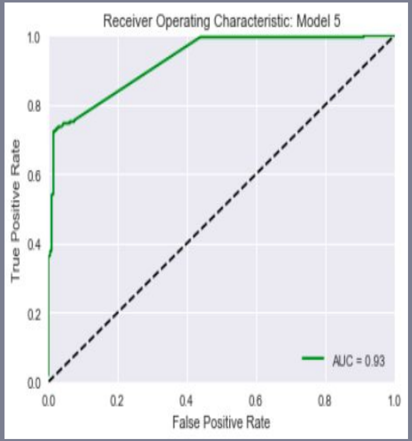
Receiver Operating Characteristic: Model 2



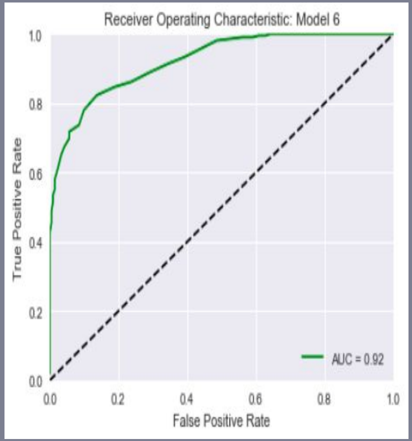
Receiver Operating Characteristic: Model 3



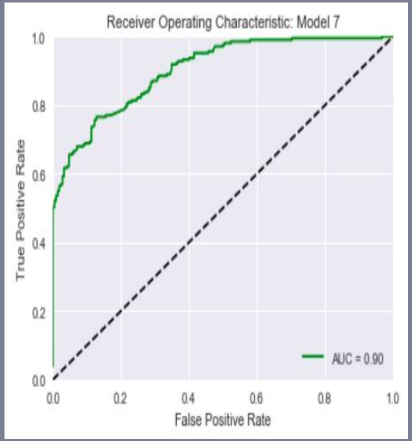
Receiver Operating Characteristic: Model 4



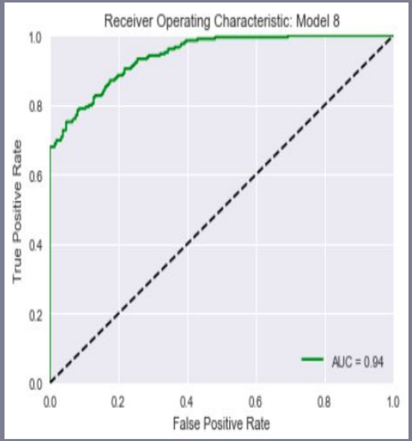
Receiver Operating Characteristic: Model 5



Receiver Operating Characteristic: Model 6



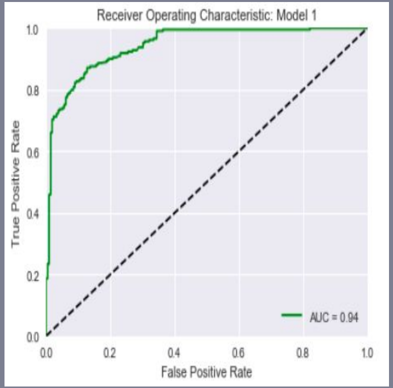
Receiver Operating Characteristic: Model 7



Receiver Operating Characteristic: Model 8

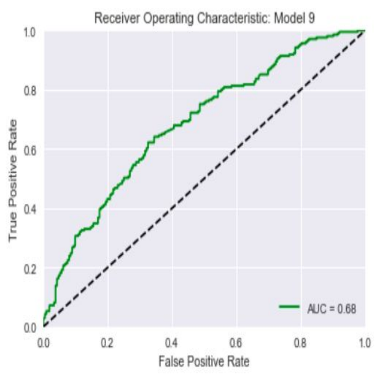
Best Model & Further Experimentation

MODEL 1



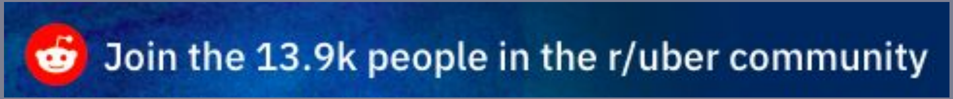
- Confusion Matrix:
  -Accuracy: 85.7%
  -Sensitivity: 77.8%
  -Specificity: 92.5%
  -Precision: 91.2%
- **Best Parameters:**
  Count Vectorizer
  -Max Features: 200
  -Min DF: 11
  -Stop Words: English
  Logistic Regression
  -C: 0.1
  -Penalty: l2 (Ridge)

MODEL 9

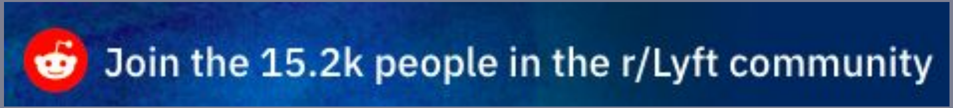


- Removed ‘Lyft’ & ‘Uber’ from Stop Words
- All other parameters identical to Model 1
- Accuracy (Baseline 50.1%)
  - **85.7% → 63.9%**
- ROC AUC Score
  - **0.94 → 0.68**

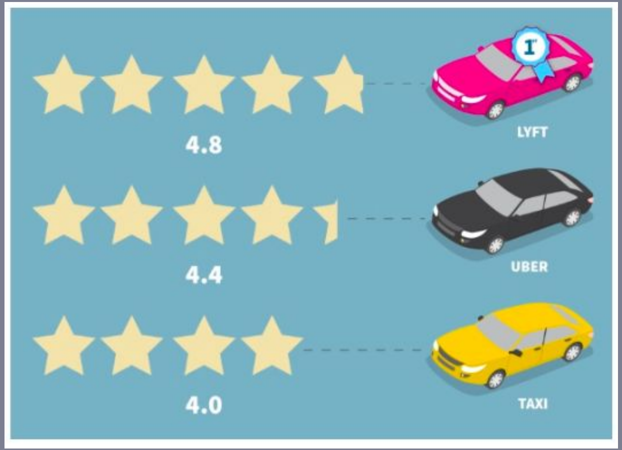
Conclusions & Recommendations



Join the 13.9k people in the r/uber community



Join the 15.2k people in the r/Lyft community



Source: http://www.businessofapps.com/data/uber-statistics

- Overall, very similar posts / almost indistinguishable without ‘lyft’ or ‘uber’ included
- Free, unfiltered feedback
  - Neutral platform
  - Prevent problems
  - Ideas for new features
  - Insight to rival company
- Internal analysis
  - Sentiment scores
  - Reading levels
  - Trace occurence of post type / top words over time

Potential Classification of Post Types

WRITER → AUDIENCE

Driver → Driver

Potential Driver → Current Drivers

Rider → Rider

Drivers → Riders

Riders → Drivers

Riders/Drivers → Company Itself

CONTEXT

Advice (alerts/warnings)

Complaints

Questions (functionality/etiquette)

Mass requests

Sharing info/links

SILICON
VALLEY
hooli
facebook
ORACLE
hp
YouTube
Google
UBER
lyft

Source: https://youtu.be/Doab-73xmqQ : Silicon Valley (HBO) - intro comparison (Season 01-04)