CSE 258

Web Mining and Recommender Systems

Assignment 2

Assignment 2

- Open-ended
- Due **Dec 3**
- Submissions should be made via gradescope

Assignment 2

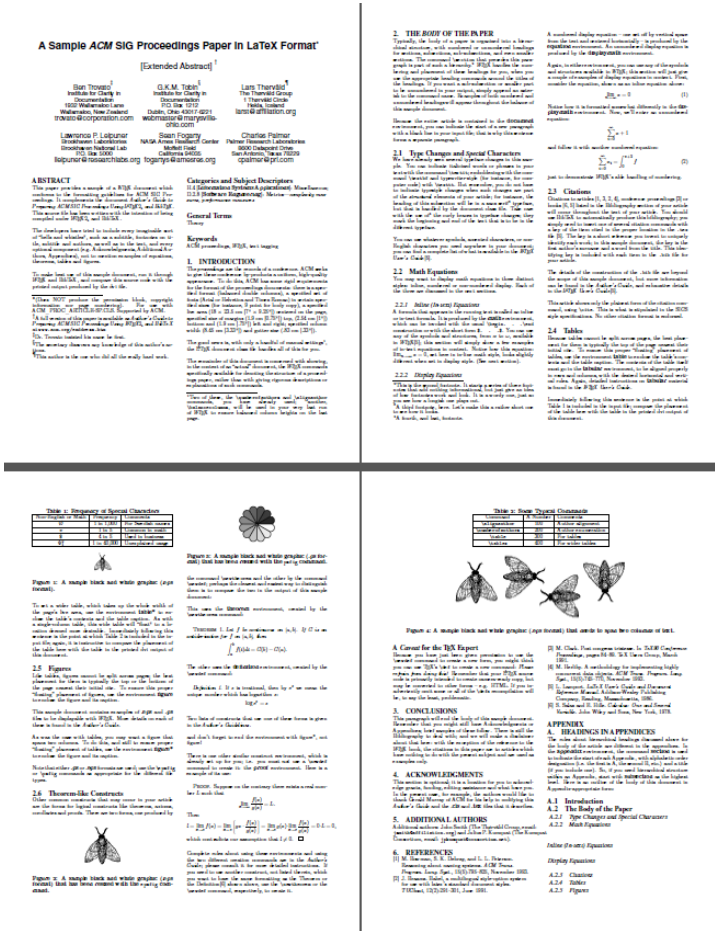
Basic tasks:

1. Identify a dataset to study and describe its basic properties
2. Identify a predictive task on this dataset and describe the features that will be relevant to it
3. Describe what model/s you will use to solve this task
4. Describe literature & research relevant to the dataset and task
5. Describe and analyze results

Assignment 2

Evaluation

- E.g. about this much:



A Sample *ACM* SIG Proceedings Paper in LaTeX Format[*]

[Extended Abstract] [†]

Ben Trovato[‡]
Institute for Clarity in Documentation
1932 Wallamaloo Lane
Wallamaloo, New Zealand
trovato@corporation.com

G.K.M. Tobin[§]
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld[¶]
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Lawrence P. Leipuner
Brookhaven Laboratories
Brookhaven National Lab
P.O. Box 5000
lleipuner@researchlabs.org

Sean Fogarty
NASA Ames Research Center
Moffett Field
California 94035
fogartys@amesres.org

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

ABSTRACT

Categories and Subject Descriptors

General Terms

Theory

Keywords

1. INTRODUCTION

2. THE *BODY* OF THE PAPER

(acm proceedings format)

https://www.acm.org/sigs/publications/proceedings-templates

Assignment 2

Teams of one to four

Assignment 2

1. **Identify a dataset to study**
   - My own repository of Recommender Systems datasets:
   - https://cseweb.ucsd.edu/~jmcauley/datasets.html

Assignment 2

1. Identify a dataset to study

- Beer data

(http://snap.stanford.edu/data/Ratebeer.txt.gz
http://snap.stanford.edu/data/Beeradvocate.txt.gz)

- Wine data

(http://snap.stanford.edu/data/cellartracker.txt.gz)

- Sensor data

(https://github.com/rpasricha/MetroInsightDataset)

Assignment 2

1. Identify a dataset to study

- Reddit submissions

(http://snap.stanford.edu/data/web-Reddit.html)

- Facebook/twitter/Google+ communities

(http://snap.stanford.edu/data/egonets-Facebook.html
http://snap.stanford.edu/data/egonets-Gplus.html
http://snap.stanford.edu/data/egonets-Twitter.html)

- Many many more from other sources, e.g.

http://snap.stanford.edu/data/

Use whatever you like, as long as it's **big**
(e.g. 50,000 datapoints minimum)

Assignment 2

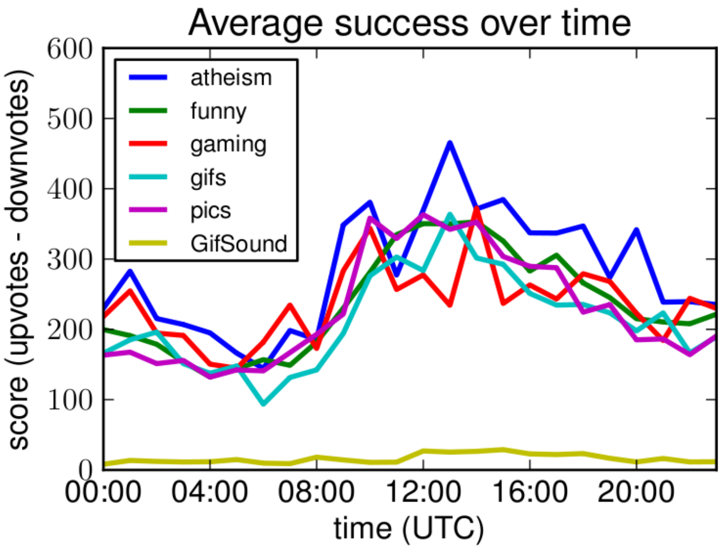
1b: Perform an **exploratory analysis** on this dataset to identify interesting phenomena

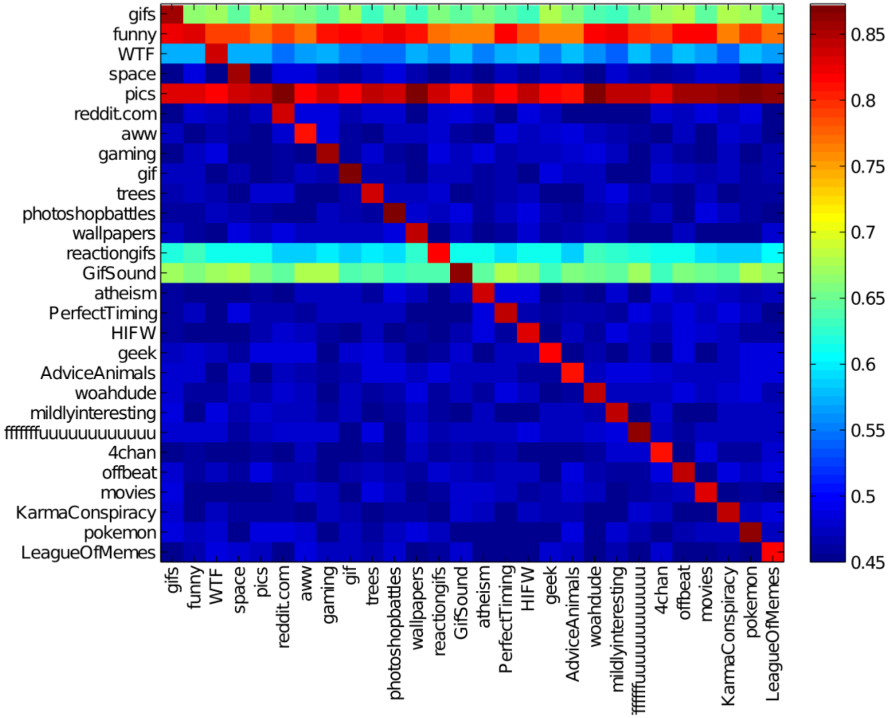
- Start with basic results, e.g. for a recommender systems type task, how many users/items/entries are there, what is the overall distribution of ratings, what time period does the dataset cover etc.

Assignment 2

1b: Perform an **exploratory analysis** of this dataset to identify interesting phenomena

e.g.





Assignment 2

2. Identify a **predictive task** on this dataset

- How will you assess the validity of your predictions and confirm that they are significant?
- Did you have to do pre-processing of your data in order to obtain useful features?
- How do the results of your exploratory analysis justify the features you have chosen?

Assignment 2

3. Select/design an appropriate model

- How will you evaluate the model? Which models from class are relevant to your predictive task, and why are other models inappropriate?
- It's totally fine here to implement a model that we covered in class, e.g. for a classification task you could implement svms+logistic regression+naïve Bayes
- You should also compare the results of different feature representations to identify which ones are effective
- What are the relevant baselines that can be compared?
- If you used a complex model, how did you optimize it?
    - What issues did you face scaling it up to the required size?
    - Any issues overfitting?
    - Any issues due to noise/missing data etc.?

Assignment 2

4. Describe related literature

- If you used an existing dataset, where did it come from and how was it used there?
- What other similar datasets have been used in the past and how?
- What are the state-of-the-art methods for the prediction task you are considering? Were you able to borrow any ideas from these works for your model? What features did they use and are you able to use the same ones?
- What were the main conclusions from the literature and how do they differ from/compare to your own findings?

Assignment 2

5. Describe your results

- Of the different models you considered, which of them worked and which of them did not?
- What is the interpretation of the parameters in your model? Which features ended up being predictive? Can you draw any interesting conclusions from the fitted parameters?

Assignment 2

Example

Maybe I want to use **restaurant data** to build a model of people's tastes in different locations

Assignment 2

1. Perform an **exploratory analysis** of this dataset to identify interesting phenomena

- How many users/items/ratings are there? Which are the most/least popular items and categories?
- What is the geographical spread of users, items, and ratings?
- Do people give higher/lower ratings to more expensive items, or items in certain countries/locations?

Assignment 2

2. Identify a **predictive task** on this dataset

- Predict what rating a person will give to a business based on the time of year, the past ratings of the user, and the geographical coordinates of the business
- Predict which businesses will succeed or fail based on its geographical location, or based on its early reviews
- What model/s and tools from class will be appropriate for this task or suitable for comparison? Are there any other tools *not* covered in class that may be appropriate?

Assignment 2

2b. Identify features that will be relevant to the task at hand

- Ratings, users, geolocations, time
- Ratings as a function of price
- Ratings as a function of location
  - How to represent location in a model? Just using a linear predictor of latitude/longitude isn’t going to work…

Assignment 2

3. Select an appropriate model

- Some kind of latent-factor model
- How to incorporate the geographical term? Should we cluster locations? Use the location as a regularizer? (etc.)
- How can we optimize this (presumably complicated) model?

Assignment 2

4. Describe related literature

- Relevant literature or predicting ratings
- Literature on using geographical features for various predictive tasks
- Literature on predicting long-term outcomes from time series data
- Literature on predicting future ratings from early reviews, herding etc.

Assignment 2

5. Describe results and conclusions

- Did features based on geographical information help? If not why not?
- Which locations are the most price sensitive according to your predictor?
- Do people prefer restaurants that are unlike anything in their area, or restaurants which are exactly the same as others in their area?

Assignment 2

Example 2

Maybe I want to use **reddit data** to see what makes submissions successful

(http://snap.stanford.edu/data/web-Reddit.html)

Assignment 2

1. Perform an **exploratory analysis** of this dataset to identify interesting phenomena

- How many users/submissions are there? How does activity differ across subreddits?
- What times of day are submissions most commented on or most rated?
- Do people give more/fewer votes to submissions that have long/short titles, or which use certain words?

Assignment 2

2. Identify a **predictive task** on this dataset

- Predict whether a post will have a large number of comments or a high rating
- Predict whether there will be a large *discrepancy* between the number of comments and the positivity of ratings a post receives
- What model/s and tools from class will be appropriate for this task or suitable for comparison? Are there any other tools *not* covered in class that may be appropriate?

Assignment 2

2b. Identify features that will be relevant to the task at hand

- Votes, users, subreddits, time
- Resubmissions of the same content & the success or failure of previous submissions
- Text of the post title

Assignment 2

3. Select an appropriate model

- Some kind of regression
- Need to use gradient descent or is there a closed-form solution?
- What are the hyperparameters and how do we regularize?
- How can you incorporate the temporal terms?

Assignment 2

4. Describe related literature

- Relevant literature or predicting votes on Reddit
- Literature on virality in social media
- Literature on using text for predictive tasks
- Literature on temporal forecasting or user preference modeling

Assignment 2

5. Describe results and conclusions

- What features helped you to predict whether content would be controversial or not?
- Does the text of the title help to predict whether a submission will be controversial or get many comments but a low vote?
- Which subreddits generate more controversial content than others?

Assignment 2

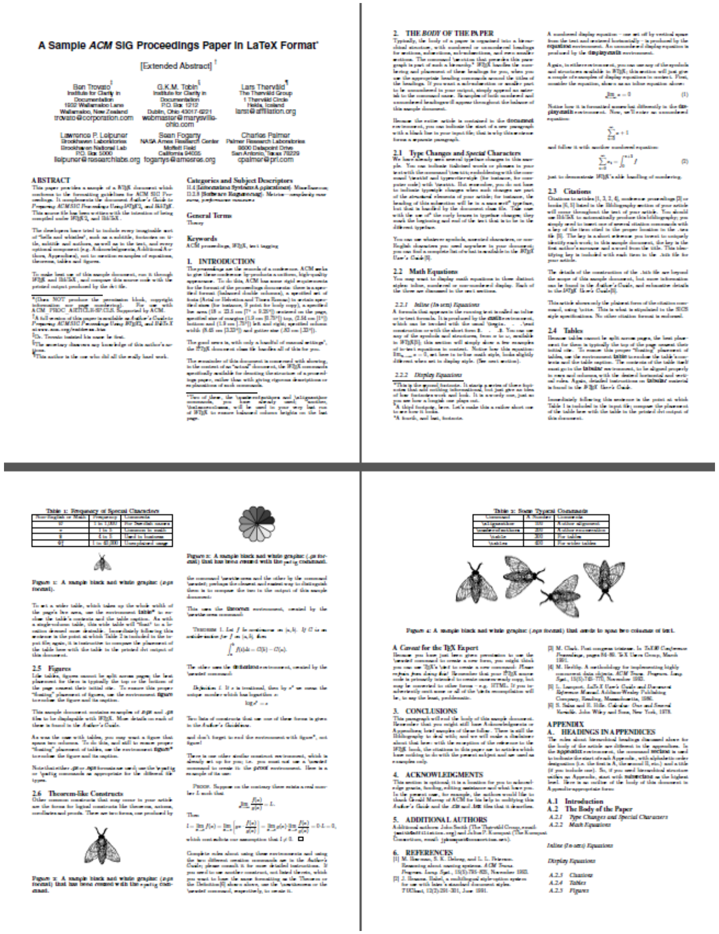
Evaluation

- These 5 sections will be worth (roughly) 5 marks each (for a total of 25% of your grade)
- Assignments can be done **in groups of up to 3 (or 4).** The marking scheme is the same regardless of group size.
- Length is not strict, but should be about 4 pages in small-font double-column format.

Assignment 2

Evaluation

- E.g. about this much:



A Sample *ACM* SIG Proceedings Paper in LaTeX Format[*]

[Extended Abstract][†]

Ben Trovato[‡]
Institute for Clarity in Documentation
1932 Wallamaloo Lane
Wallamaloo, New Zealand
trovato@corporation.com

G.K.M. Tobin[§]
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld[¶]
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Lawrence P. Leipuner
Brookhaven Laboratories
Brookhaven National Lab
P.O. Box 5000
lleipuner@researchlabs.org

Sean Fogarty
NASA Ames Research Center
Moffett Field
California 94035
fogartys@amesres.org

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

ABSTRACT

Categories and Subject Descriptors

General Terms

Theory

Keywords

1. INTRODUCTION

2. THE BODY OF THE PAPER

(acm proceedings format)

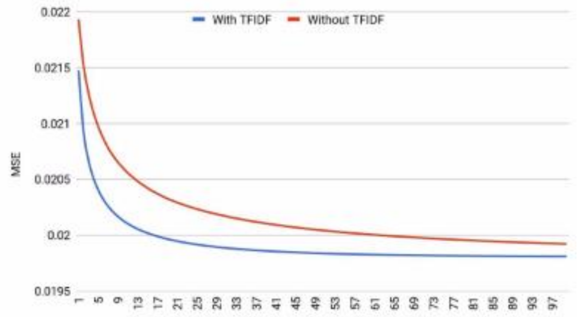
https://www.acm.org/sigs/publications/proceedings-templates

Data Mining and Predictive Analytics

Assignment 2 – examples of previous assignments

Supervised funniness detection in the New Yorker cartoon caption contest

"I was just transferred to the fraternity ward."



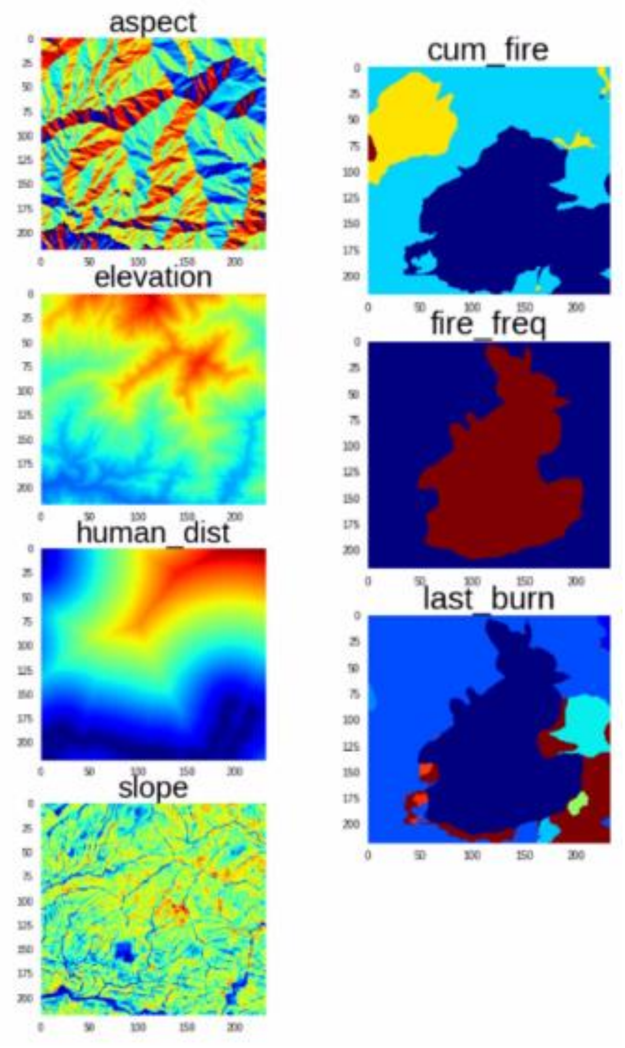
TF-IDF vs non-TF-IDF models

- Predict whether a caption will be scored as "funny" by human judges
- 65 images, 320k captions
- Scores from 1.0 – 2.75

- BoW methods w/ and w/o TF-IDF
- Dimensionality-reduction-based feature representations

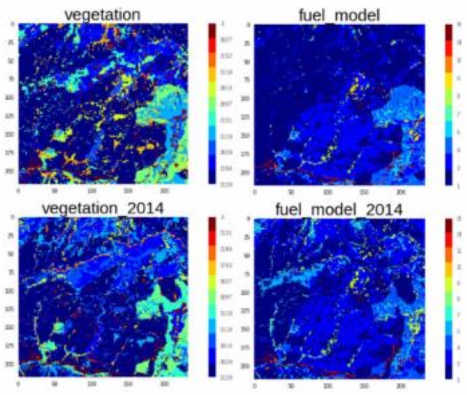
Melissa Wright

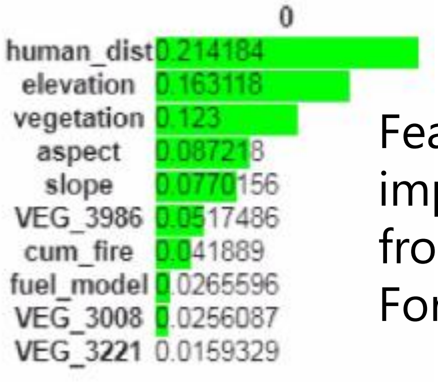
Predicting Vegetation Changes as Responses to Forest Fires



- Geological data from LANDFIRE program and FRAP (Fire and Resource Assessment Program), 1992-2012
- Estimate changes as a result of forest fires

$$y = x_{2012\ vegetation} == x_{2014\ vegetation}\ \forall x \in X$$





	0
human_dist	0.214184
elevation	0.163118
vegetation	0.123
aspect	0.087218
slope	0.0770156
VEG_3986	0.0517486
cum_fire	0.041889
fuel_model	0.0265596
VEG_3008	0.0256087
VEG_3221	0.0159329

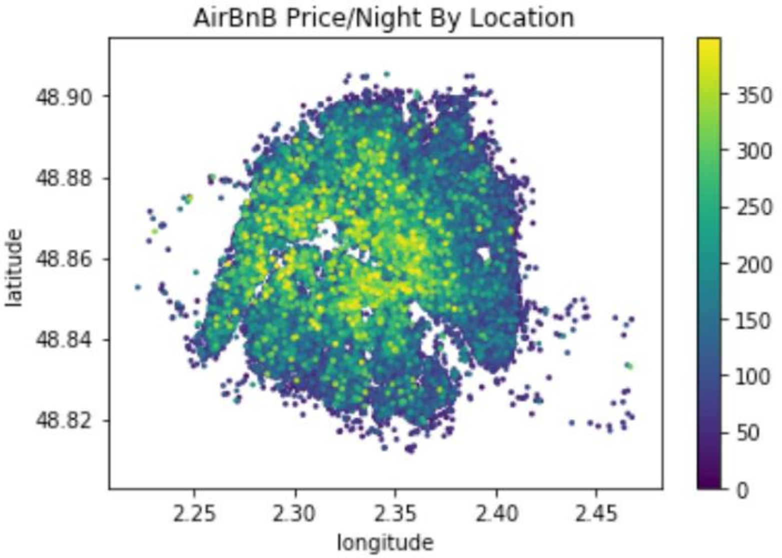
Feature importance from Random Forest Model

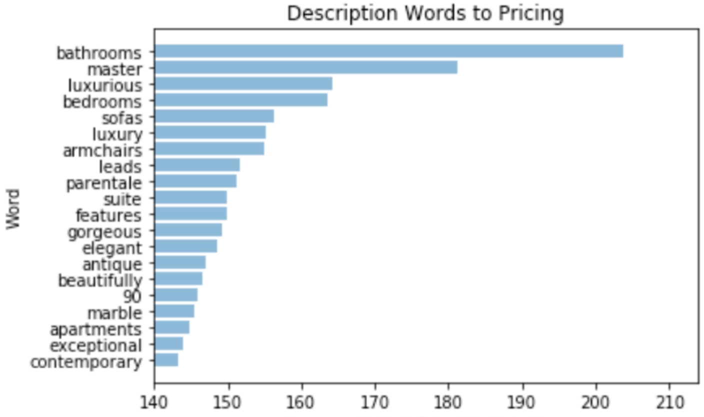
Tony Salim

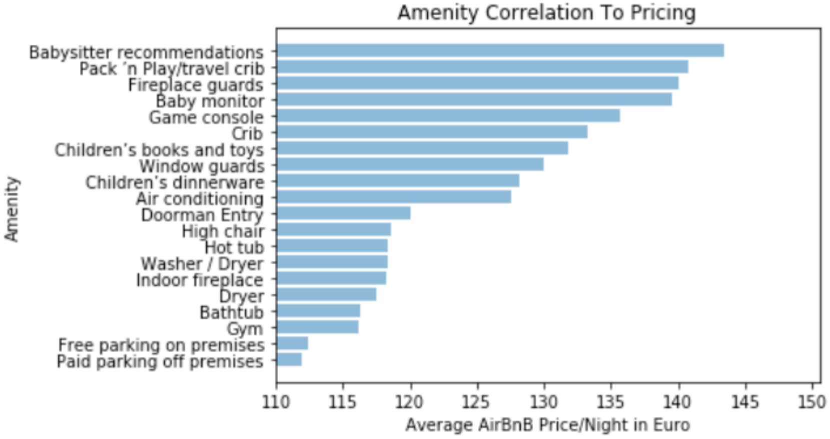
AirBnB Price Per Night Prediction

Price Range	€ 0.00 to € 7,790.00
Mean	€ 96.12
Median	€ 75.00
Standard Deviation	€ 99.30

- AirBnB Paris data
- Predict listing price given various features



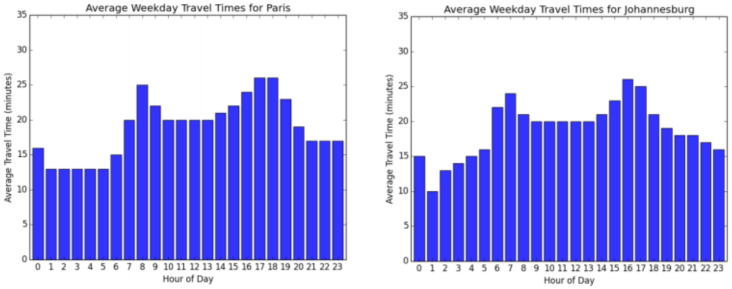




Peter Mai

Uber Everywhere: Exploring Movement

Feature	Description
Hour of day (hod)	Simple hour of the day feature.
Source ID	Simple source ID feature.
Destination ID	Simple destination ID feature.
Hour of day historical mean*	Mean travel category of trips for this hour of day.
Source ID historical mean*	Mean travel category of trips from this source ID.
Destination ID historical mean*	Mean travel category of trips from this destination ID.
Source-Destination ID pair historical mean*	Mean travel category of trips from specific source ID-destination ID pair.



Weekday travel times in two cities

- Anonymized Uber Movement data from 7 cities
- Trip time given source, destination, and hour

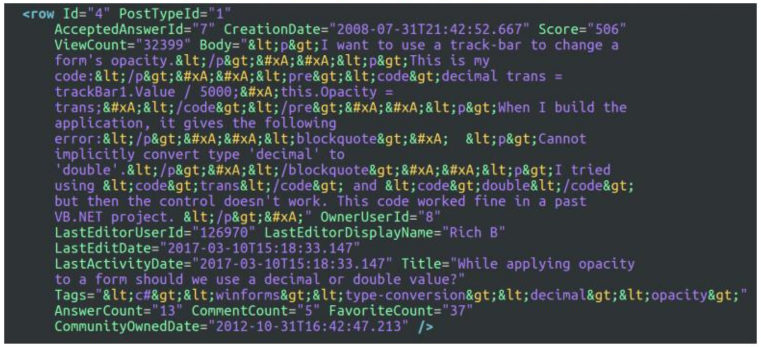
Feature Representation	Week Category	Results
hod, source ID, dest ID	Weekday	26.544%
	Weekend	29.247%
hod mean, source ID mean, dest ID mean	Weekday	26.788%
	Weekend	29.113%
hod, source ID, dest ID, hod mean, source ID mean, dest ID mean, combined source ID-dest ID mean	Weekday	21.318%
	Weekend	25.024%
hod, combined source ID-dest ID mean	Weekday	79.218% / 79.975%*
	Weekend	87.041% / 87.146%*

SVM,
Random Forest
MLP

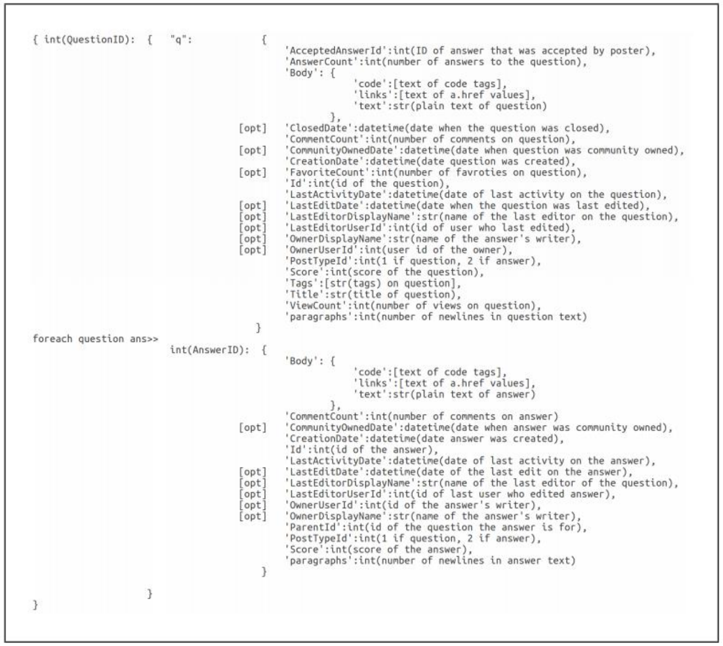
Tynan Dewes, David Thomson

Predicting the Accepted Answer for StackOverflow Questions

Figure 1: Example Entry in Posts.xml



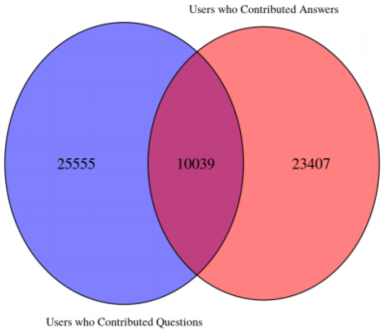
```
<row Id="4" PostTypeId="1"
    AcceptedAnswerId="7" CreationDate="2008-07-31T21:42:52.667" Score="506"
    ViewCount="32399" Body="&lt;p&gt;I want to use a track-bar to change a
    form's opacity.&lt;/p&gt;&#xA;&#xA;&lt;p&gt;This is my
    code:&lt;/p&gt;&#xA;&#xA;&lt;pre&gt;&lt;code&gt;decimal trans =
    trackBar1.Value / 5000;&#xA;this.Opacity =
    trans;&#xA;&lt;/code&gt;&lt;/pre&gt;&#xA;&#xA;&lt;p&gt;When I build the
    application, it gives the following
    error:&lt;/p&gt;&#xA;&#xA;&lt;blockquote&gt;&#xA;  &lt;p&gt;Cannot
    implicitly convert type 'decimal' to
    'double'.&lt;/p&gt;&#xA;&lt;/blockquote&gt;&#xA;&#xA;&lt;p&gt;I tried
    using &lt;code&gt;trans&lt;/code&gt; and &lt;code&gt;double&lt;/code&gt;
    but then the control doesn't work. This code worked fine in a past
    VB.NET project. &lt;/p&gt;&#xA;" OwnerUserId="8"
    LastEditorUserId="126970" LastEditorDisplayName="Rich B"
    LastEditDate="2017-03-10T15:18:33.147"
    LastActivityDate="2017-03-10T15:18:33.147" Title="While applying opacity
    to a form should we use a decimal or double value?"
    Tags="&lt;c#&gt;&lt;winforms&gt;&lt;type-conversion&gt;&lt;decimal&gt;&lt;opacity&gt;"
    AnswerCount="13" CommentCount="5" FavoriteCount="37"
    CommunityOwnedDate="2012-10-31T16:42:47.213" />
```

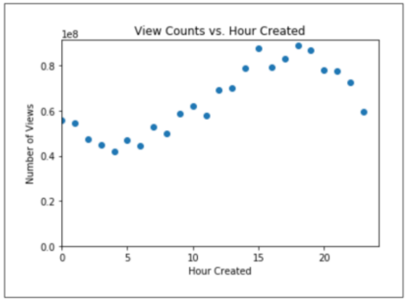


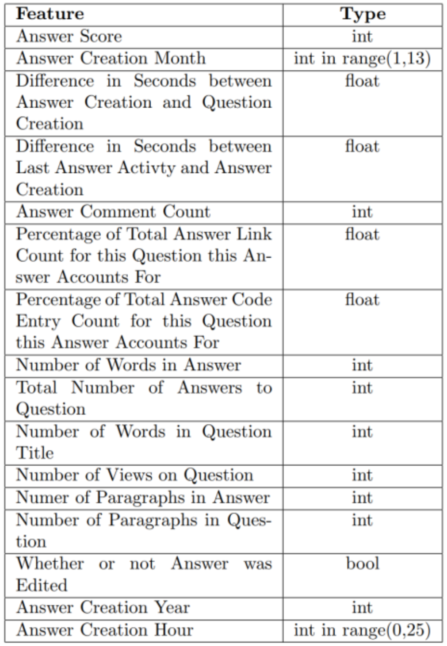
```
{ int(QuestionID):  {   "q":            {
                                                'AcceptedAnswerId':int(ID of answer that was accepted by poster),
                                                'AnswerCount':int(number of answers to the question),
                                                'Body': {
                                                            'code':[text of code tags],
                                                            'links':[text of a.href values],
                                                            'text':str(plain text of question)
                                                        },
                                        [opt]   'ClosedDate':datetime(date when the question was closed),
                                                'CommentCount':int(number of comments on question),
                                        [opt]   'CommunityOwnedDate':datetime(date when question was community owned),
                                                'CreationDate':datetime(date question was created),
                                        [opt]   'FavoriteCount':int(number of favroties on question),
                                                'Id':int(id of the question),
                                                'LastActivityDate':datetime(date of last activity on the question),
                                        [opt]   'LastEditDate':datetime(date when the question was last edited),
                                        [opt]   'LastEditorDisplayName':str(name of the last editor on the question),
                                        [opt]   'LastEditorUserId':int(id of user who last edited),
                                        [opt]   'OwnerDisplayName':str(name of the answer's writer),
                                        [opt]   'OwnerUserId':int(user id of the owner),
                                                'PostTypeId':int(1 if question, 2 if answer),
                                                'Score':int(score of the question),
                                                'Tags':[str(tags) on question],
                                                'Title':str(title of question),
                                                'ViewCount':int(number of views on question),
                                                'paragraphs':int(number of newlines in question text)
                                        }
foreach question ans>>
                        int(AnswerID):  {
                                                'Body': {
                                                            'code':[text of code tags],
                                                            'links':[text of a.href values],
                                                            'text':str(plain text of answer)
                                                        },
                                                'CommentCount':int(number of comments on answer)
                                        [opt]   'CommunityOwnedDate':datetime(date when answer was community owned),
                                                'CreationDate':datetime(date answer was created),
                                                'Id':int(id of the answer),
                                                'LastActivityDate':datetime(date of last activity on the answer),
                                        [opt]   'LastEditDate':datetime(date of the last edit on the answer),
                                        [opt]   'LastEditorDisplayName':str(name of the last editor of the question),
                                        [opt]   'LastEditorUserId':int(id of last user who edited answer),
                                        [opt]   'OwnerUserId':int(id of the answer's writer),
                                        [opt]   'OwnerDisplayName':str(name of the answer's writer),
                                                'ParentId':int(id of the question the answer is for),
                                                'PostTypeId':int(1 if question, 2 if answer),
                                                'Score':int(score of the answer),
                                                'paragraphs':int(number of newlines in answer text)
                                        }

                    }
}
```

- Large dataset of StackOverflow posts
- Predict which answer is marked as “accepted” (classification)


Users who Contributed Answers
25555
10039
23407
Users who Contributed Questions





Feature	Type
Answer Score	int
Answer Creation Month	int in range(1,13)
Difference in Seconds between Answer Creation and Question Creation	float
Difference in Seconds between Last Answer Activty and Answer Creation	float
Answer Comment Count	int
Percentage of Total Answer Link Count for this Question this Answer Accounts For	float
Percentage of Total Answer Code Entry Count for this Question this Answer Accounts For	float
Number of Words in Answer	int
Total Number of Answers to Question	int
Number of Words in Question Title	int
Number of Views on Question	int
Numer of Paragraphs in Answer	int
Number of Paragraphs in Question	int
Whether or not Answer was Edited	bool
Answer Creation Year	int
Answer Creation Hour	int in range(0,25)

Mustafa Guler, Jessica Kwok, Joseph Thomas

Bitcoin Price Prediction using ARIMA, Linear Regression and Deep Learning

- Does historical Bitcoin data contain enough information to predict its future value (“autoregression”-like task)

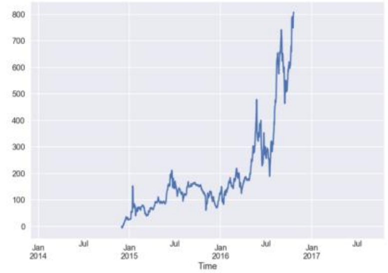


Fig. 4. Percentage Return on Investment in 1 year

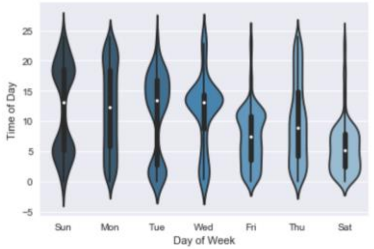
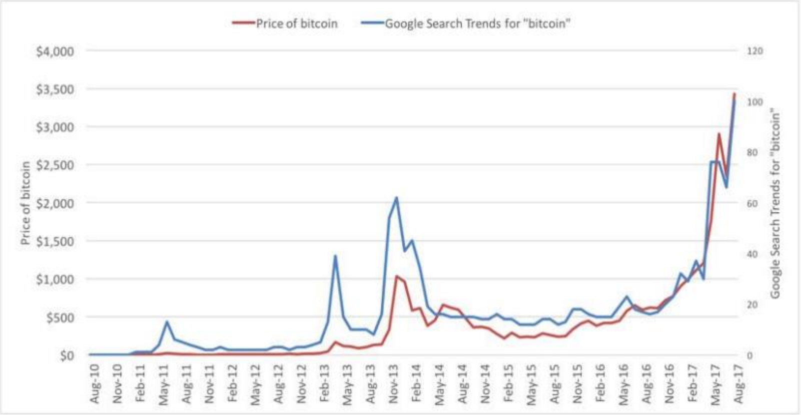


Fig. 6. Violin plot describing best time of day to invest in bitcoin



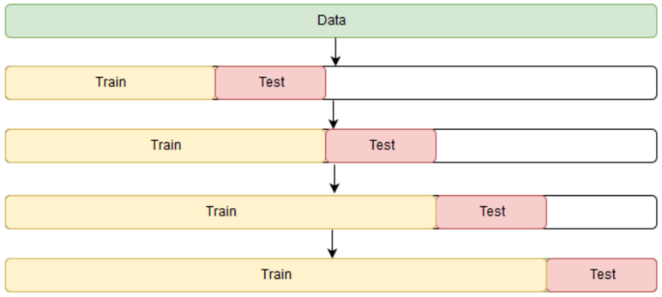
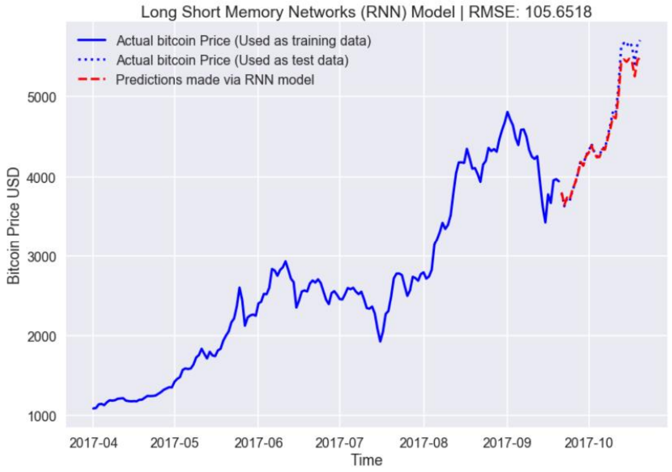


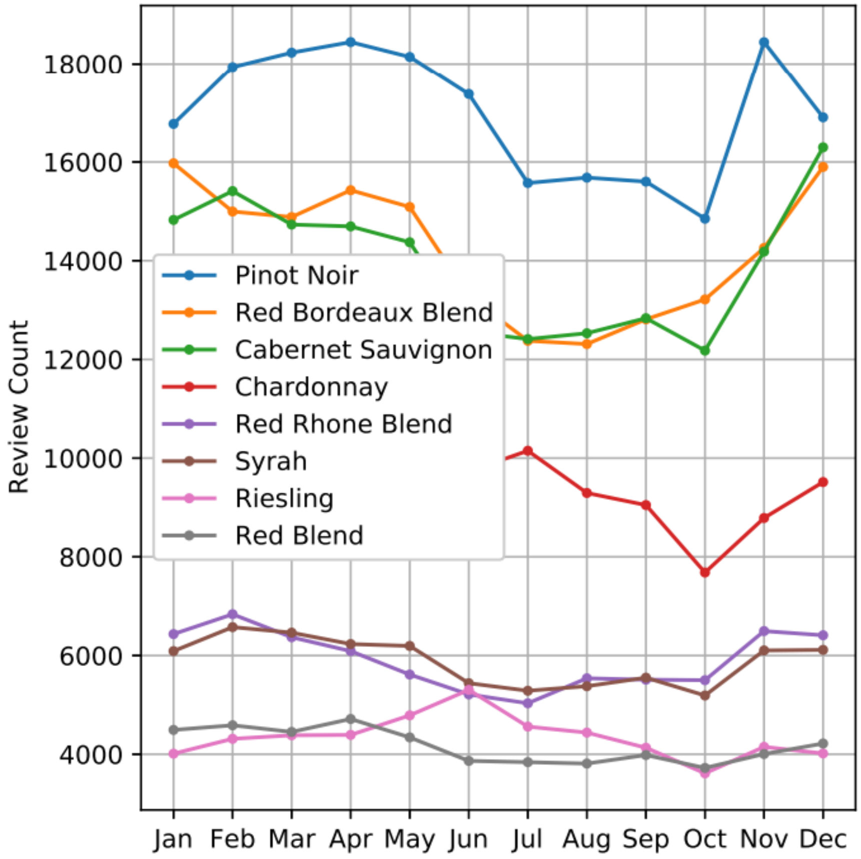
Fig. 7. Cross Validation on a rolling basis [10]



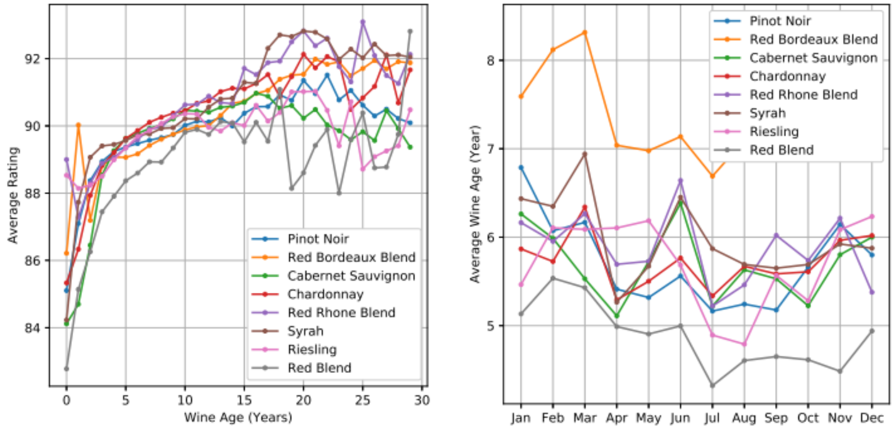
Evaluation Metric	Trained Time Series Models			
	Baseline	***ARIMA***	***Linear Regression***	***LSTM***
RSS	8,529,112	8,148,537	629,980	334,868
MSE	284,303	271,617	20,999	11,162
RMSE	533.20	521.16	144.91	**105.65**

Aman Aggarwal, Gurkanwal Singh Batra

Predicting Wine Popularity Using Temporal Features



- Wine demand appears to exhibit seasonal variability. Can this be predicted?

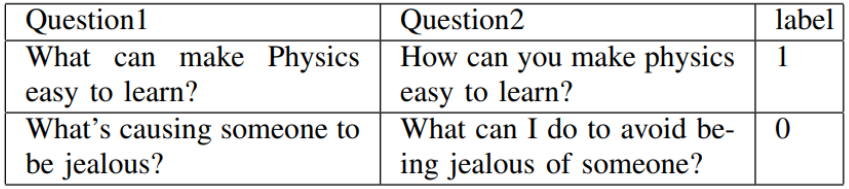


consumption of “high quality” wine is seasonal

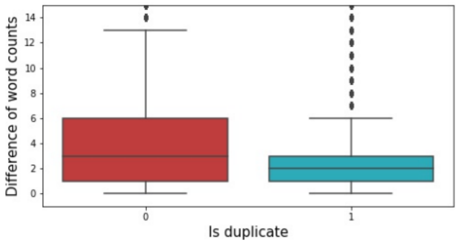
prediction	accuracy
random selection	0.25
pick most popular	0.714
k-nearest neighbor	0.786

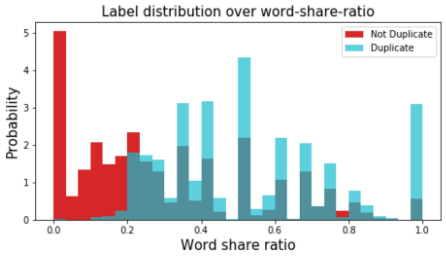
Canruo Ying

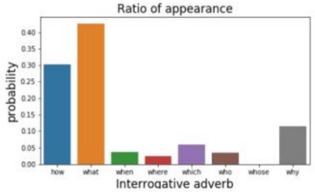
Duplicate Question Detection on Quora

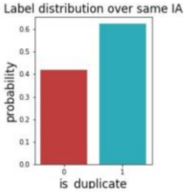


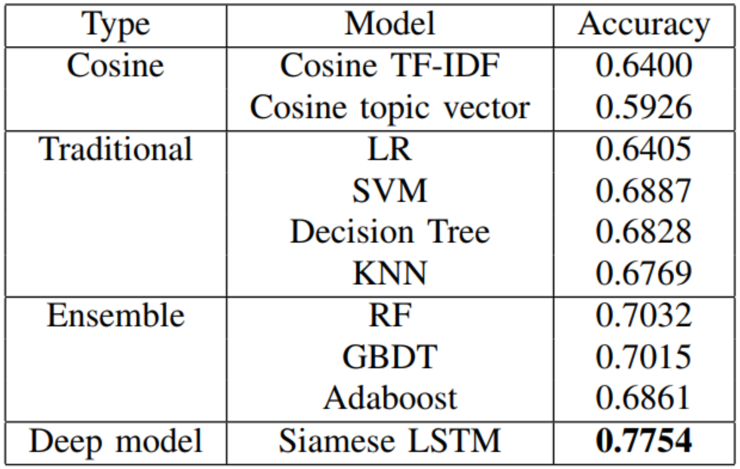
Question1	Question2	label
What can make Physics easy to learn?	How can you make physics easy to learn?	1
What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	0





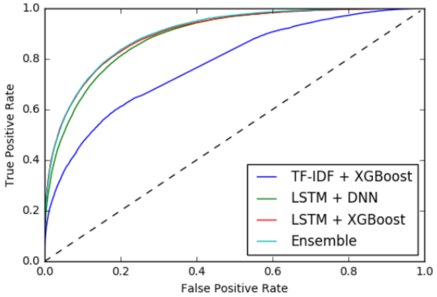






Type	Model	Accuracy
Cosine	Cosine TF-IDF	0.6400
	Cosine topic vector	0.5926
Traditional	LR	0.6405
	SVM	0.6887
	Decision Tree	0.6828
	KNN	0.6769
Ensemble	RF	0.7032
	GBDT	0.7015
	Adaboost	0.6861
Deep model	Siamese LSTM	**0.7754**

Yi Luo,
Jingtao Song,
Haoting Chen



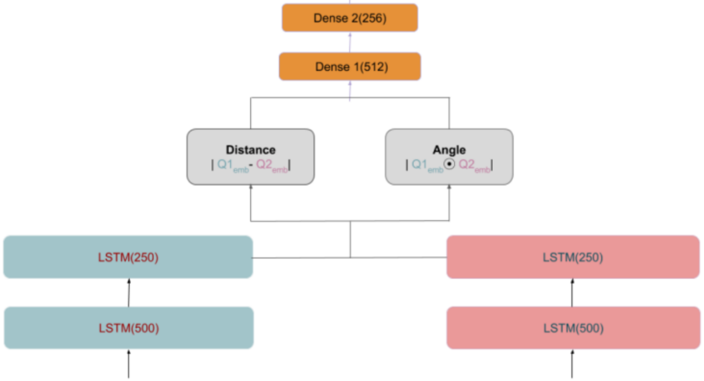
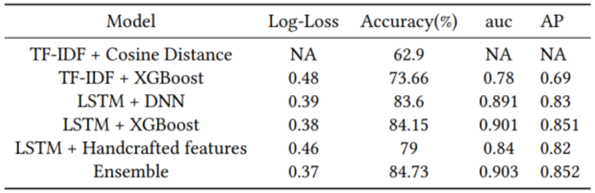


Figure 5: LSTM-based feature extractor followed by handcrafted feature extraction

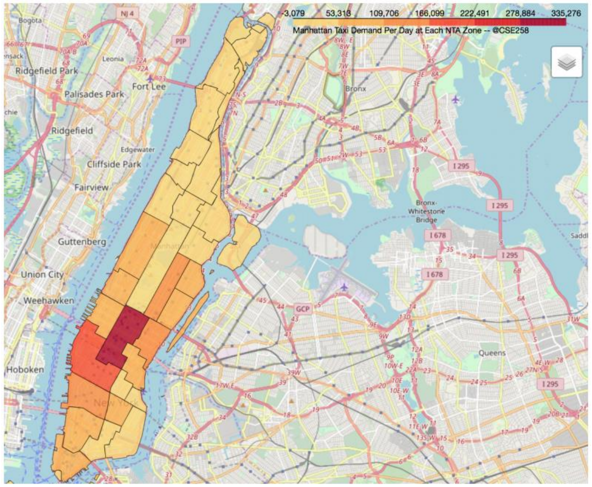
Table 2: Comparative evaluation of all models

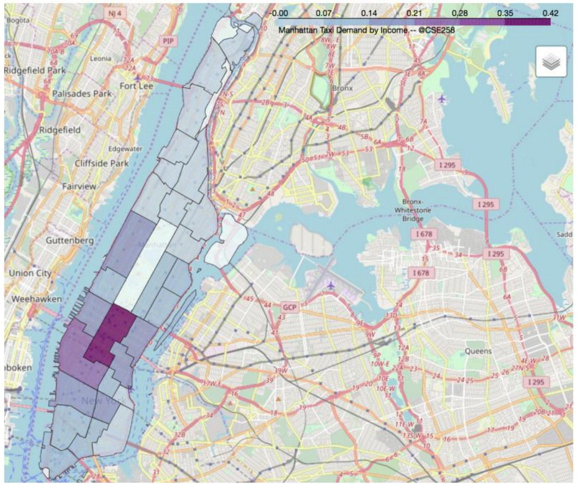


Model	Log-Loss	Accuracy(%)	auc	AP
TF-IDF + Cosine Distance	NA	62.9	NA	NA
TF-IDF + XGBoost	0.48	73.66	0.78	0.69
LSTM + DNN	0.39	83.6	0.891	0.83
LSTM + XGBoost	0.38	84.15	0.901	0.851
LSTM + Handcrafted features	0.46	79	0.84	0.82
Ensemble	0.37	84.73	0.903	0.852

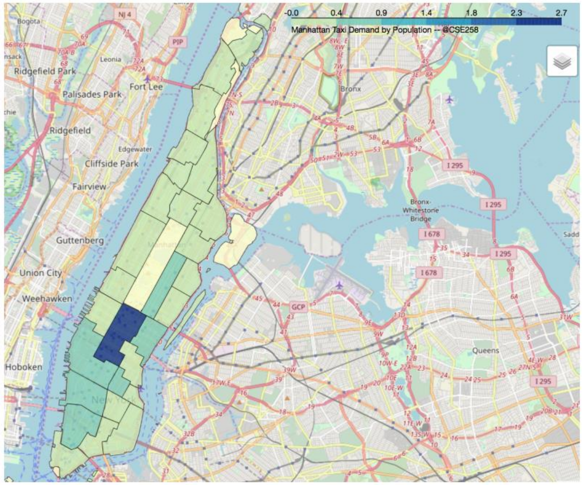
Vaibhav Gandhi, Akshaya
Purohit, Aditya Verma

NYC Taxi Demand Prediction

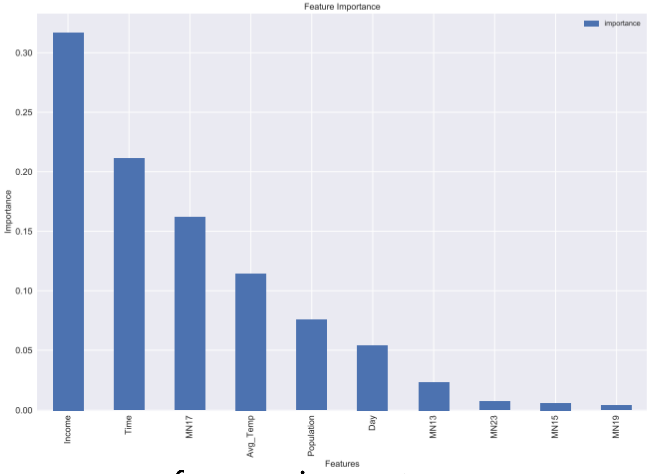




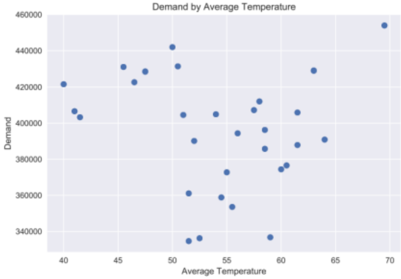
income



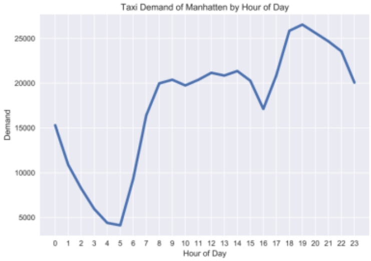
population



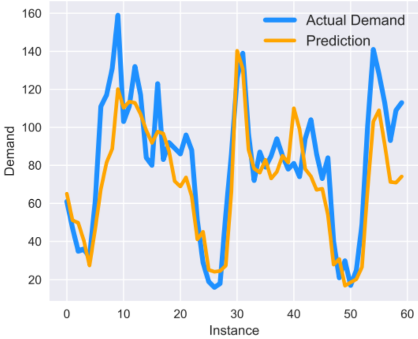
feature importance
(gradient boosted decision tree)



temperature

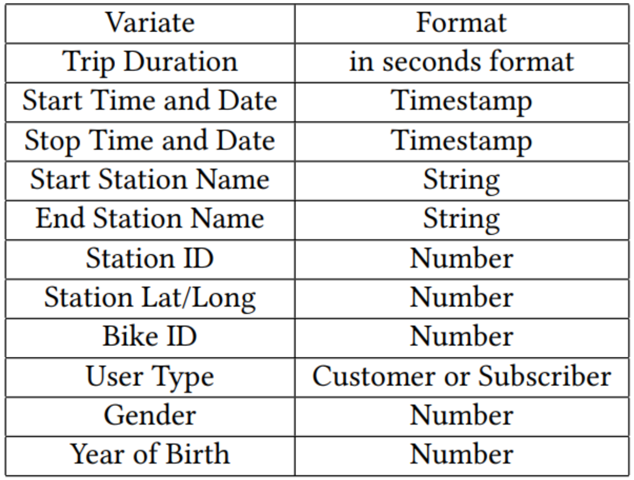


hour



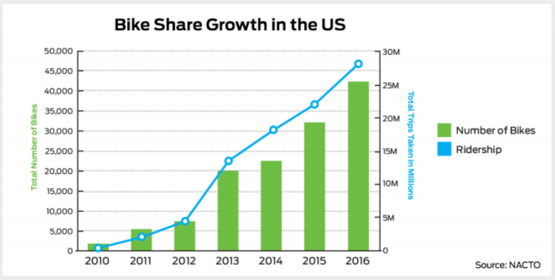
Siyu Jiang, Simran Kapur, Siddharth Dinesh

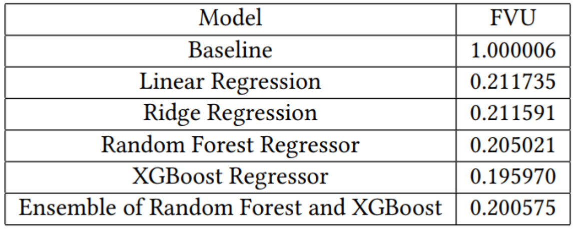
NYC Bike Trip Duration Prediction



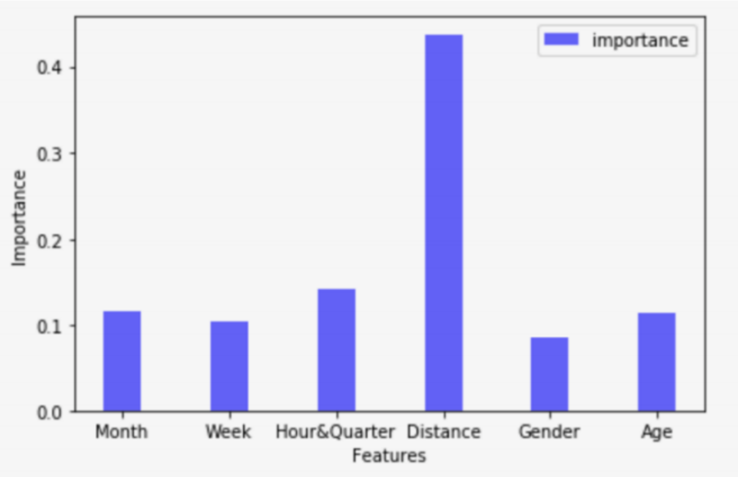
Variate	Format
Trip Duration	in seconds format
Start Time and Date	Timestamp
Stop Time and Date	Timestamp
Start Station Name	String
End Station Name	String
Station ID	Number
Station Lat/Long	Number
Bike ID	Number
User Type	Customer or Subscriber
Gender	Number
Year of Birth	Number

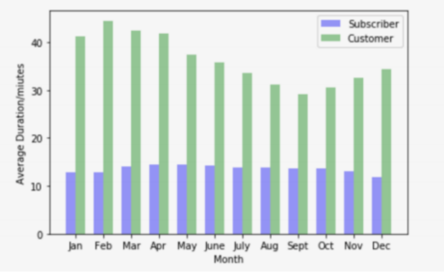
North Bergen
West New York
MANHATTAN
Weehawken
Central Park Zoo
HELL'S KITCHEN
Times Square
ASTORIA
MIDTOWN
Empire State Building
Hoboken
LONG ISLAND CITY
SUNNYSIDE
Washington Square Park
GREENPOINT
NEWPORT
LOWER MANHATTAN
New York
Brooklyn Bridge
WILLIAMSBURG
BROOKLYN HEIGHTS
BUSHWICK
BEDFORD-STUYVESANT
RED HOOK
BROOKLYN
NEW JERSEY
NEW YORK



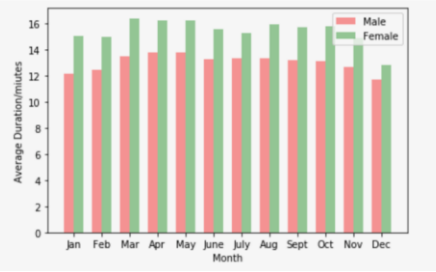


Model	FVU
Baseline	1.000006
Linear Regression	0.211735
Ridge Regression	0.211591
Random Forest Regressor	0.205021
XGBoost Regressor	0.195970
Ensemble of Random Forest and XGBoost	0.200575





subscriber vs. customer

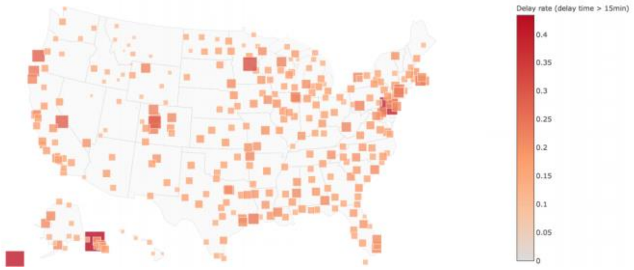


duration vs. gender

Zhuo Cheng, Tianran Zhang, Jiamin He

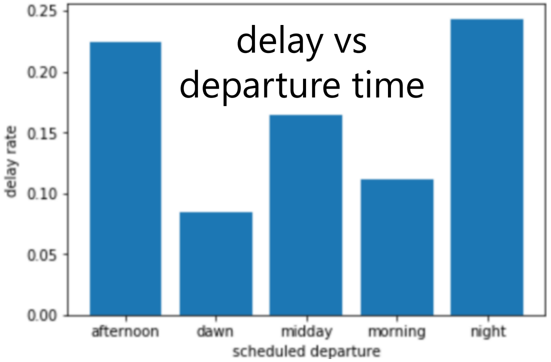
Airline Delay Prediction

delay vs origin



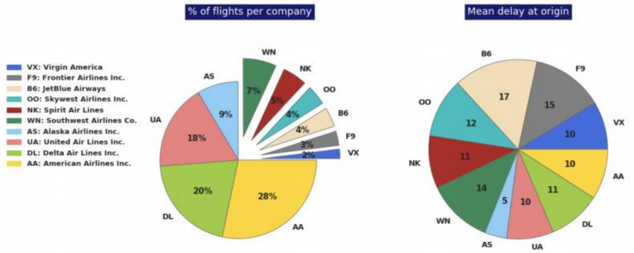
delay vs route

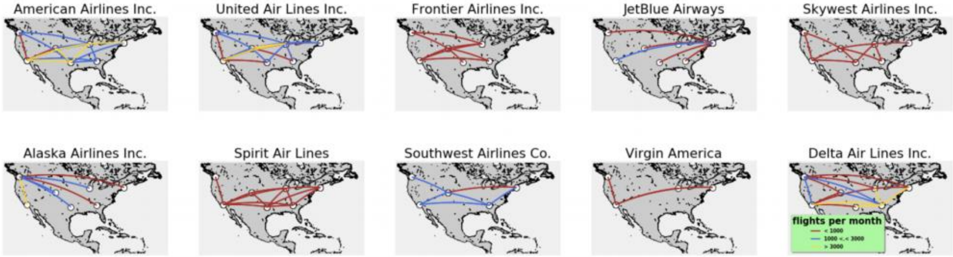
delay vs departure time



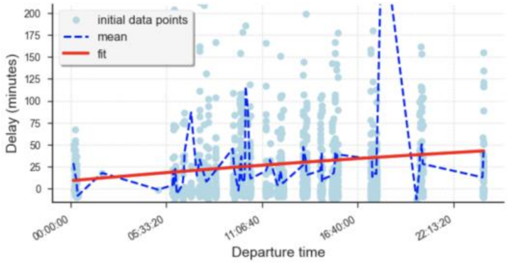
Methods	AUC scores	Precision	Recall	F_1 score	Accuracy
Baseline	0	0	0	0	0.798
Naive Bayes	0.6294	0.3049	0.4044	0.3467	0.6920
Logistic Regression	0.6492	0.3478	0.34	0.3367	0.7345
Random Forest	0.6129	0.2441	0.0074	0.0140	0.7975
Neural Network	0.6404	0.5218	0.0677	0.1150	0.7946

Ran Wang
Qianlong Qu
Yuan Qi
Zijia Chen





Feature Name	Encoding	Dimension
airline	one-hot	10
scheduled_departure	one-hot	24
month	one-hot	12
day_of_month	one-hot	31
day_of_week	one-hot	7
origin_airport	one-hot	7
destination_airport	one-hot	7
distance	float	1
wind_speed	float	1
visibility_in_miles	float	1
sky_coverage	one-hot	5



KNN, SVM, Softmax regression

Qian Zhang
Simeng Zhu
Feng Jiang
He Qin

Questions?