Project Portfolio

Lena Choe (lenachoe2024@u.northwestern.edu)

Contents

- Predictive Models using Machine Learning
    - Real Estate Rent Prediction
    - TBM (tunnel boring machine) Utility Prediction
- Kpop Lyric Text Analysis
- US News University Ranking Survey Data Mining
- BASIC256 Programming Tutorial

Prediction – Real Estate Price

Built a model predicting monthly rent for commercial offices. Utilized location, office space, age of the building, nearby subway stations, and other amenities such as free parking.

Applied various modeling techniques such as linear regression, decision tree, and neural networks, and choose the best model by comparing performances of the models.

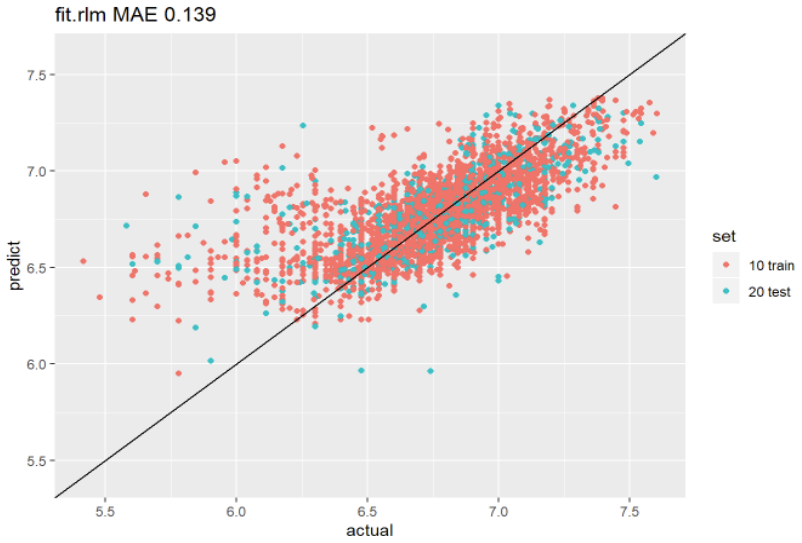
Created a Shiny app summarizing results from various models and comparing performances.

Robust Linear Regression

Linear regression but it removes outliers that may alter the equation too much, to provide a better and more accurate fit.

```
fit.rlm <- MASS::rlm(form, data = df0)
pred.rlm <- predict(fit.rlm, newdata = test_df)
```

```
show_res_all(fit.rlm)
```



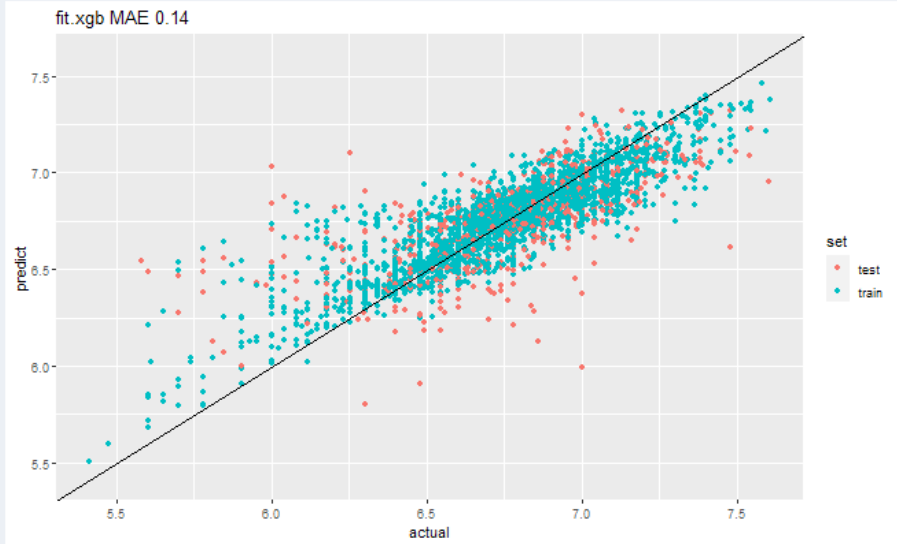

```
summary(fit.rlm)
```

```
##
## Call: rlm(formula = form, data = df0)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.224689 -0.098409  0.002111  0.099325  0.778102
```

Result of Robust Linear Regression, created from a RMarkdown document.
Green: train samples, Red: test samples

Prediction – Real Estate Price (continued)

The one with the best result (lowest MAE error) was the model trained with XGBoost.



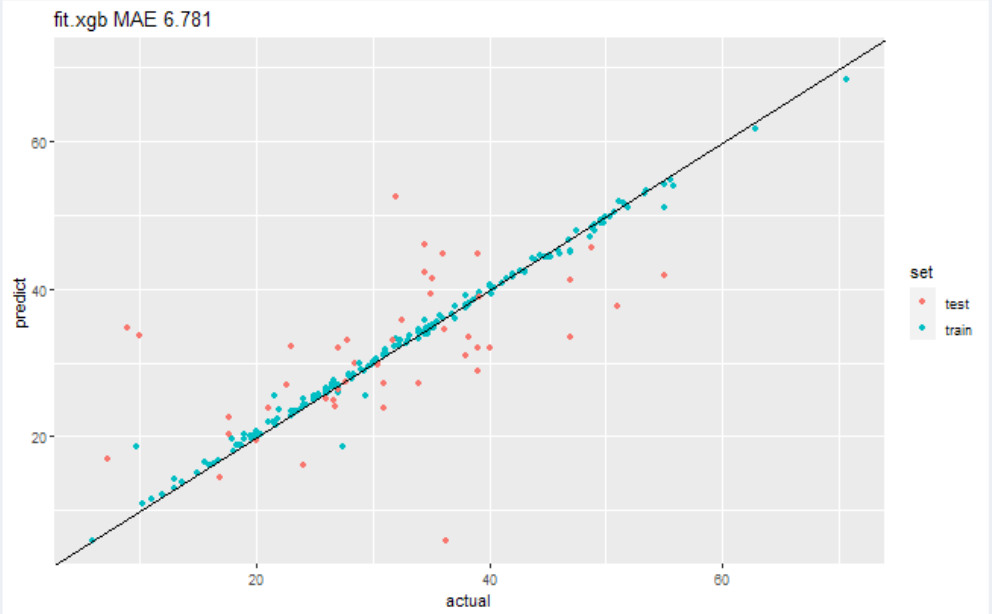

```
              Length Class              Mode
handle             1 xgb.Booster.handle externalptr
raw           656935 -none-             raw
niter              1 -none-             numeric
call               5 -none-             call
params             8 -none-             list
callbacks          1 -none-             list
feature_names     45 -none-             character
nfeatures          1 -none-             numeric
xNames            45 -none-             character
problemType        1 -none-             character
tuneValue          7 data.frame         list
obsLevels          1 -none-             logical
param              0 -none-             list
```

Prediction – Tunnel Boring Machine Utility

TBM (https://en.wikipedia.org/wiki/Tunnel_boring_machine) as it stands, is a type of construction machine used for boarding underground tunnels. We built a prediction model given (only a few) input variables of interest - boarding machine type, tunnel diameter, tunnel length in meter, and soil condition, predicting utility which is defined as the ratio of actual working days against total rental days.

Basically the same script used for rent prediction has been re-used in this case, with minor modification.

In this case, XGB also gives the best result. However, the presence of the



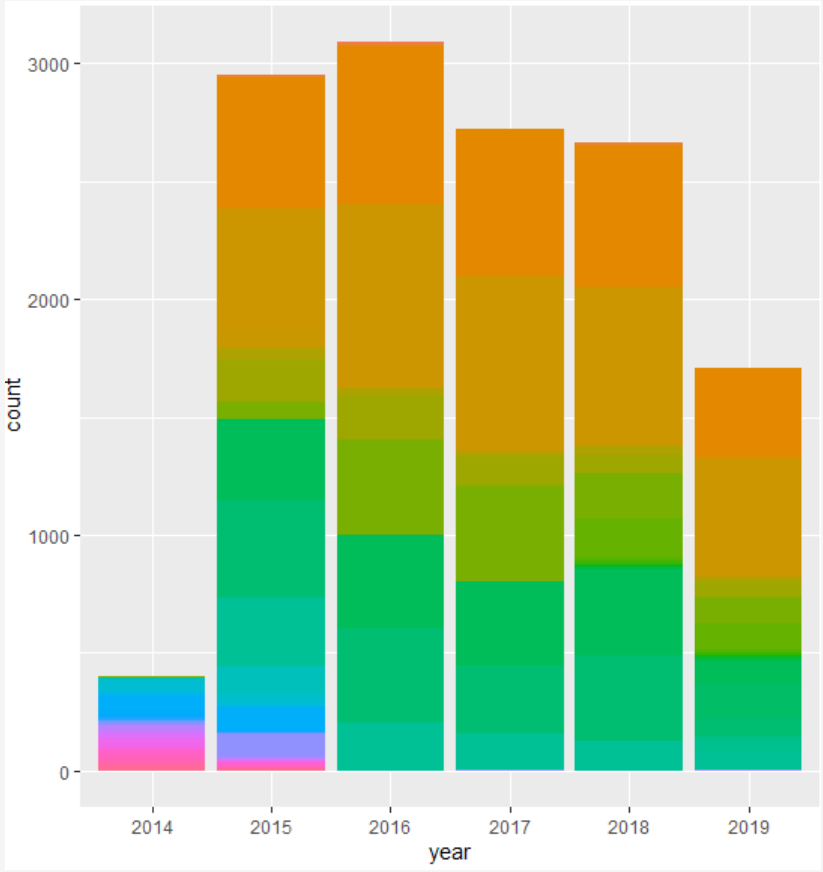
	Length	Class	Mode
handle	1	xgb.Booster.handle	externalptr
raw	479023	-none-	raw
niter	1	-none-	numeric
call	5	-none-	call
params	8	-none-	list
callbacks	1	-none-	list
feature_names	9	-none-	character
nfeatures	1	-none-	numeric
xNames	9	-none-	character
problemType	1	-none-	character
tuneValue	7	data.frame	list
obsLevels	1	-none-	logical
param	0	-none-	list

Kpop Lyric Text Analysis

Performed basic text analysis using Kpop lyrics (Korean and English translation) downloaded from https://klyrics.net/.

Did word frequency plot, investigated most frequently used words and trends in the lyrics by comparing word frequencies along the years.

I have downloaded articles from klyrics.net page, a total of 13,133 songs from 3,289 artists spanning to mid-2019.



Number of songs downloaded per year

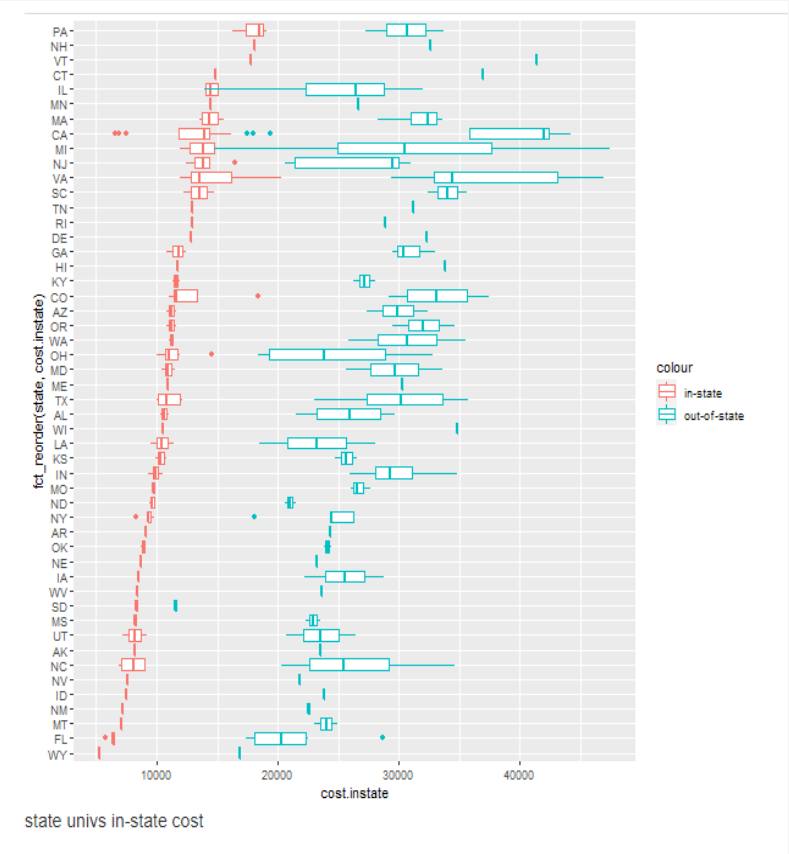
rnk	Dance	Ballad	`Rap
<int>	<chr>	<chr>	<chr>
51	girl	니가	너를
52	거야	함께	한
53	know	없어	know
54	your	baby	건
55	너무	싶어	왜
56	지금	잘	just
57	봐	않아	go
58	na	이제	내게
59	in	때	니
60	있어	있는	니가
61	좀	건	girl
62	we	보고	우리
63	한	사람	hangul
64	be	지금	싶어
65	and	참	all
66	u	사랑	너무
67	싶어	없이	다시
68	것	안	wanna
69	돼	우린	결
70	게	못	지금
71	모든	정말	우린
72	just	돼	그냥
73	now	두	못
74	같아	모두	좀
75	안	밤	what
76	do	속에	거야
77	hangul	맘	get
78	없는	곁에	네가
79	ah	할	that
80	같은	아직	잘
81	맘	the	없는
82	good	사랑해	것
83	wanna	그냥	너와
84	ooh	너에게	is
85	다시	그렇게	않아
86	come	눈을	이제
87	boy	니	때
88	마	그런	봐
89	get	혼자	with
90	what	너는	got

Word frequency rank for three genres of Dance, Ballard and Rap. Definitely Dance and Rap songs have more English words than Ballard songs.

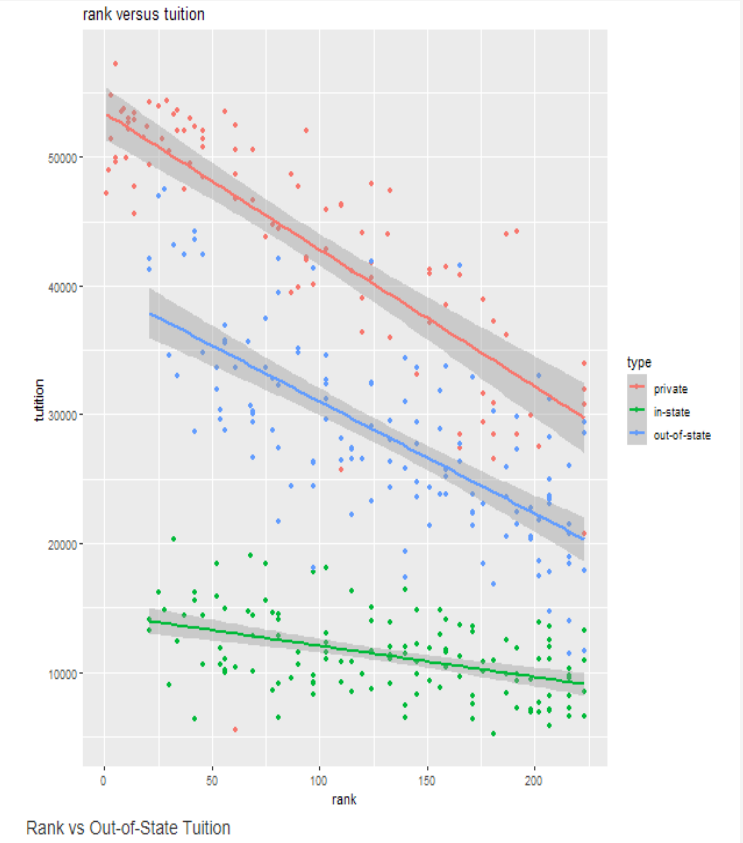
US News University Ranking Survey Data Mining

Collected college data from the "US News University Rankings" site by web scraping using R's httr package.

Post-process downloaded articles to extract information to create an R data file. Also created a shiny app to present reports showing various statistics such as tuition and rankings.



Tuition differences for in-state and out-of-state students



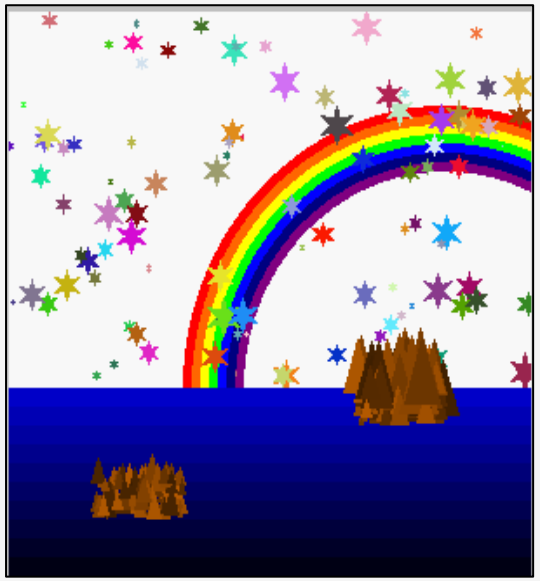
Rank versus tuition – pay $100 more for one rank up

BASIC Programming Tutorial

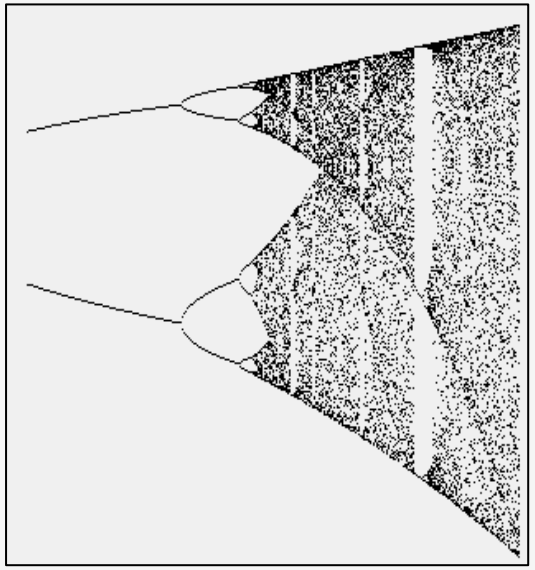
Practiced computer programming using a dialect of BASIC language (BASIC256).

Learned fundamental programming skills such as variables, control structures, algorithms, and data structures. Built simple games, graphics, and simulations.

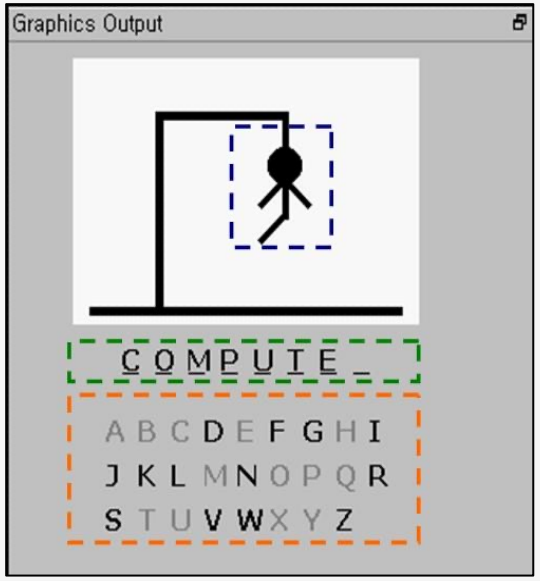
Drafted a tutorial, or a 80+page mini-book showing how to program BASIC.



A virtual landscape using geometric objects



Logistic map showing Chaos



Hangman Game