Project 3: Predicting Subreddit using Natural Language Processing

By Tenzin Wangdu

SKI VS SNOWBOARD

Subreddits:

r/ski

- 12.3k members
- “Dude, it needs to be winter faster!”
- Extracted 1000 posts
- Created in Aug 26, 2008

r/snowboard

- 1.6k members
- “for industry insiders”
- Extracted 400 posts
- Created Jan 11, 2010

Problem Statement

How can we use a predictive model to see which subreddit a post came from?

Data Collection

- Scrape from pushshift’s API
- Acquired the data using request library(res = requests.get(url, params)
- Format into a dictionary using json(data = res.json())

EDA and Cleaning

- fill all null value with ‘’
- Combine the title and selftext into one column for the target variable
- Delete all duplicates by using .drop_dulpicates()
- Convert ski and snowboard into binary column

Tokenizing and Lemmatization

RegExp Tokenizer: (‘[a-z]\w+’): it return only the lowercase letter without any punctuation or special letter.

Lemmatize: To normalize the text without any derived words.

hi guys,

finally decided to commit to a new pair of skis. have been thinking rossignol experience 84 as i have skied on multiple occasions. i'm 178cm, weight only 65kg, more of an intermediate skier (depends on how you defines it...) and mainly ski in australia, so mostly on front side, firm to slightly soft snow. the shop that i rent fro m only have 163cm and that's what i've been ski on, very versatile, fun at short crave turns but does not go very fast and chatter at high speed. perhaps it's too sho rt? the guy at the shop suggested that the length appropriate for me given my weight. i'm a decent ice skater and thus very comfortable putting weight on edges despit e being "intermediate" skier.

the other one that i've tried recently was dynastar speedzone 4x4 82 at 171cm, i found that very stable at wider turns and at speed, kind of like that feeling, but i cannot do short turns on this thing gracefully. i'm guessing i cannot put enough force on the edge unless i'm going fast. either my skills (more likely), or my weight that i cannot push down on a longer/stiffer ski at slow speed?

i'm looking for skis that able to help me grow in terms of skills, able to do short-medium crave turns and stable when i want to go fast. would appreciate your opinio n and experience!

currently i'm considering:

rossignol experience 84, 168cm

head v10, 163 or 170cm?

volkl deacon 84, 167cm

or whatever you think i should be considering! ski advice for a light weight skier

---

hi guy finally decided to commit to new pair of ski have been thinking rossignol experience a have skied on multiple occasion cm weight only kg more of an intermediat e skier depends on how you defines it and mainly ski in australia so mostly on front side firm to slightly soft snow the shop that rent from only have cm and that wha t ve been ski on very versatile fun at short crave turn but doe not go very fast and chatter at high speed perhaps it too short the guy at the shop suggested that the length appropriate for me given my weight decent ice skater and thus very comfortable putting weight on edge despite being intermediate skier the other one that ve tr ied recently wa dynastar speedzone x4 at cm found that very stable at wider turn and at speed kind of like that feeling but cannot do short turn on this thing gracefu lly guessing cannot put enough force on the edge unless going fast either my skill more likely or my weight that cannot push down on longer stiffer ski at slow speed looking for ski that able to help me grow in term of skill able to do short medium crave turn and stable when want to go fast would appreciate your opinion and experi ence amp x200b currently considering rossignol experience cm head v10 or cm volkl deacon cm or whatever you think should be considering ski advice for light weight sk ier

Model Selection

Logistic Regression with Count Vectorizer

- Remove common english by setting a stopword
- Easier to interpret the coef_
- Built pipeline and GridSearch

Random Forest with Tfidif

- Decision Tree
- stop_words, ngram_range
- Built pipeline and GridSearch

Model Performance

Model	Train Score	Test Score	Cross Val Score
CountVectorizer/ Logistic Regression	92.18	87.1	85.9
Tfidif/Random Forest	98.6	88.2	87.9

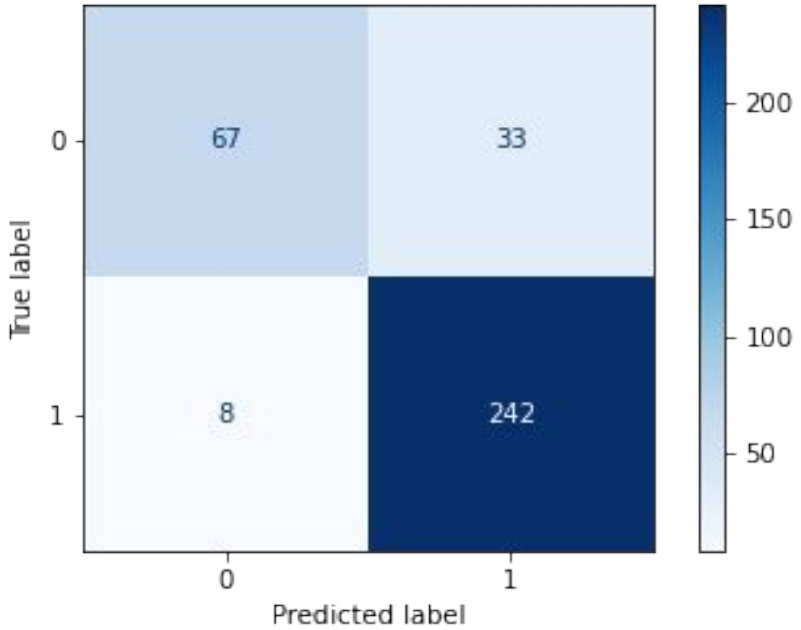
Baseline: 0.70
Best params(CountVec):

- Max_features: 700
- Min_df: 2
- Max_df : 0.4

Bestparams(Tfidf):

- Max_features: 1000
- Min_df: 3
- Max_df: 0.5

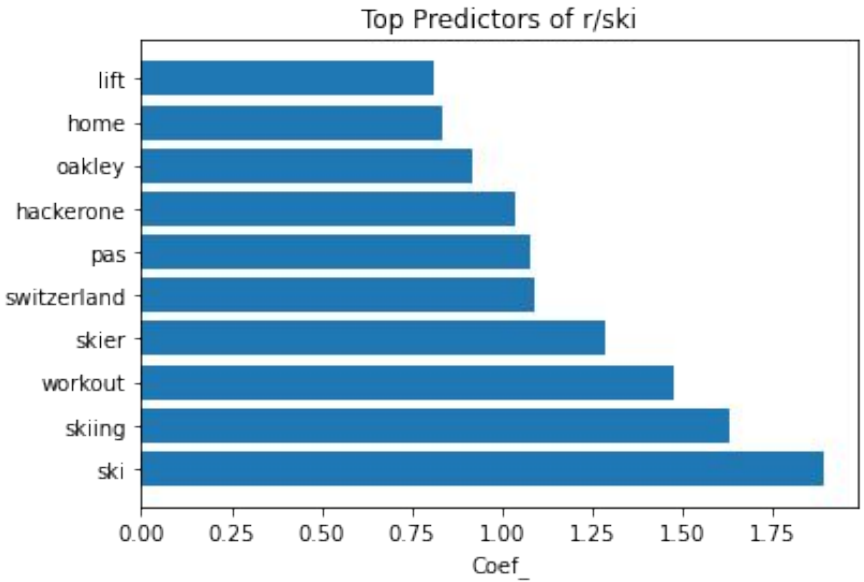
Confusion Matrix on Logistic Regression

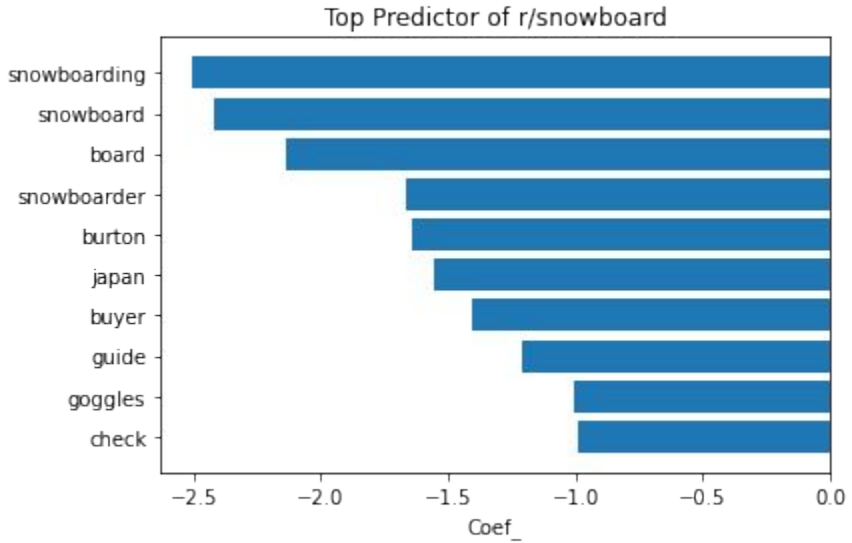


0: Snowboard, 1: ski

- Accuracy Rate of 0.88
- Misclassification: 0.117

Top Predictor for r/ski and r/snowboard





Conclusion

- Both model were successful, they both beat the baseline by more than 20%
- Logistic Regression was the better model and its coefficient makes the data more understandable
- The Next Steps would be gather new data and implement into my model to see the score

Question? Or Comments!

Most Common Words in Both Subreddit