Natural Language Processing and Classification

Using Subreddit Communities to Optimize Engagement with Consumers

Farah Malik
Project 3
June 2023

Problem Statement

Background

Participation in **winter sports** has been rising over the last decade, with a **26% increase** in the active number of snowboarders and skiers.[1] It is only projected to increase, with high consumer values on outdoor activity, experiences, and health. For the first time, the interest in snowboarding has outpaced that of skiing.[2]

Client and Issue

Burton Snowboards, Inc. is a leader in the winter sports space, but is facing increasing competition in a niche market and among evolving consumer interests and demands.

It is critical that Burton finds way to **expand its customer based and establish consumer loyalty.**

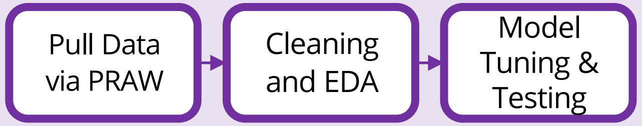
Objective

The objective of this project is to develop a classification model that **predicts the source of a Reddit post** based on its text.

With this, we will **identify opportunities** for Burton that will enhance its market share, generate financial opportunities, and optimize products and marketing.

Methodology & Analysis

- Acquire data via Python Reddit API Wrapper (**PRAW**)
- Conduct **exploratory data analysis**, cleaning, data transformation via Vectorizing
- Build and test various **classification models**
- Evaluate with **accuracy report**


Pull Data via PRAW → Cleaning and EDA → Model Tuning & Testing

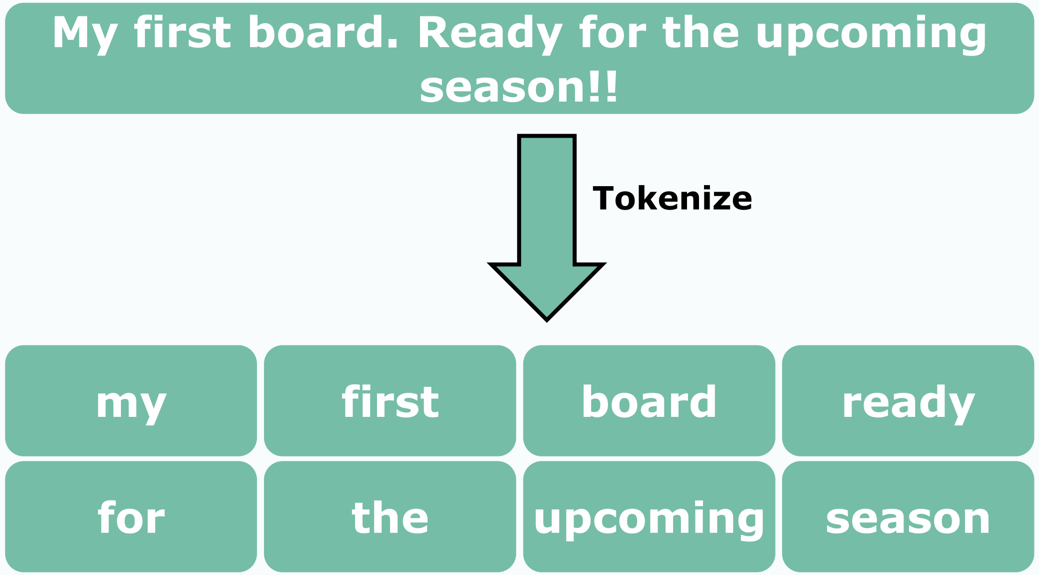
[1] https://kenver.com/blogs/news/state-snow-sports/
[2] https://www.factmr.com/report/229/snowboard-equipment-market

Pre-Processing and Transformation

Baseline Model

Skiers: 51.2%
Snowboarders: 48.8%

RegexpTokenizer



My first board. Ready for the upcoming season!!

Tokenize

Text Transformation – CountVectorizer and TfidfVectorizer

Board Brand
Burton
Burton
Avalanche
K2
Avalanche

Burton	Avalanche	K2
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Modeling, Tuning, and Testing

Logistic Regression

- **Rationale:** Binary classification, large amount of data for sufficient performance, interpretability
- **Hyperparameters Tested**: C, penalty
- **Precision Outcome**: 91.33%

Multinomial Naïve Bayes

- **Rationale:** Many discrete features (e.g., word counts for text classification), performance with textual data, computational efficiency and scalability
- **Hyperparameters Tested**: Alpha
- **Precision Outcome**: 87.94%

Support Vector Machine

- **Rationale:** High performance for text classification task, ability to handle non-linear relationships, versatility in kernel choices
- **Hyperparameters Tested**: C, kernel, degree
- **Precision Outcome**: 91.95%

Modeling, Tuning, and Testing (Cont'd)

Random Forest

- **Rationale:** Reduce overfitting, performance with textual data, capture of feature importances to aid interpretability
- **Hyperparameters Tested**: N_estimators, max_depth
- **Precision Outcome**: Accuracy not high enough, did not continue to calculating precision

Extra Trees

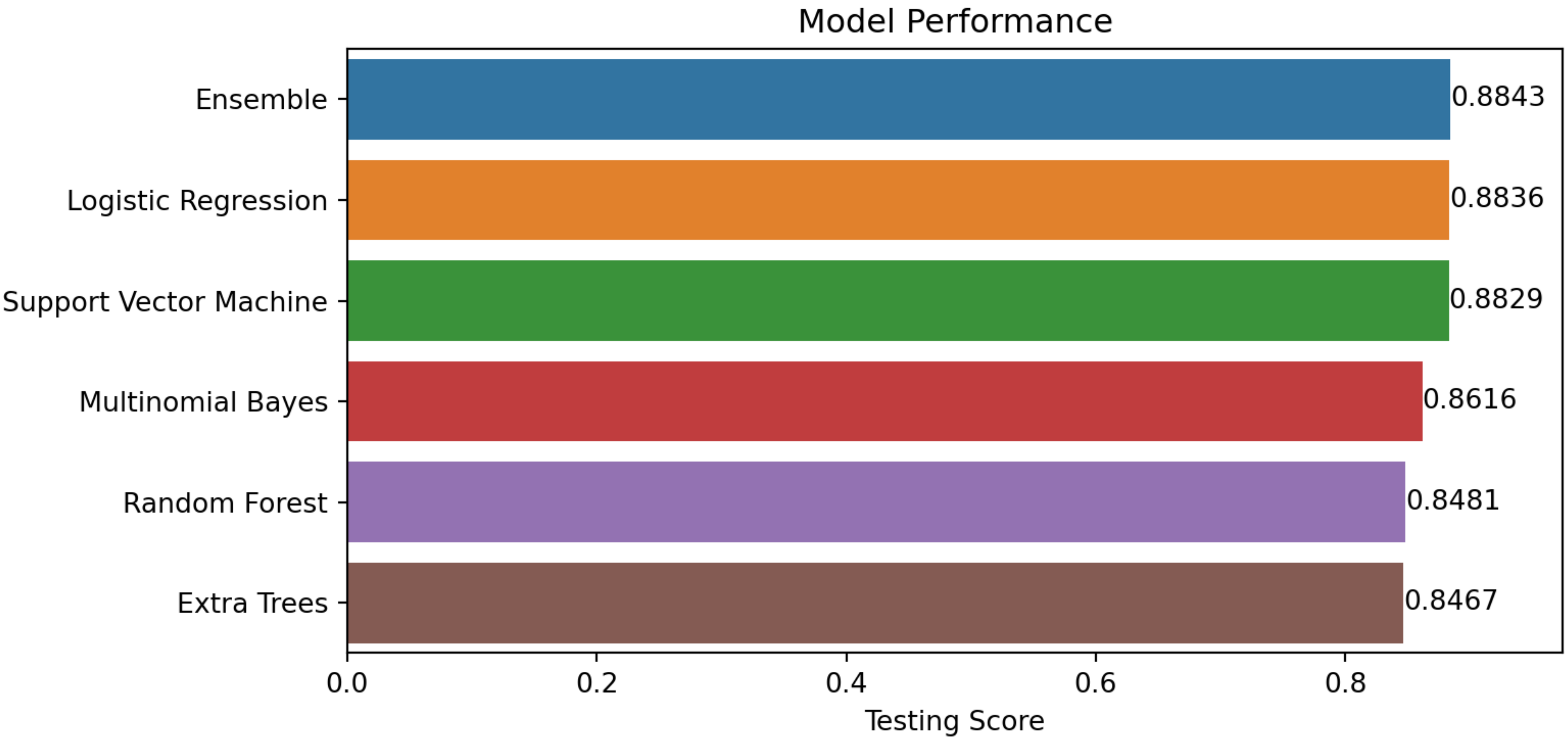
- **Rationale:** Reduce overfitting, extra randomness to further reduce variance, performance with textual data, capture of feature importances, computational efficiency
- **Hyperparameters Tested**: N_estimators, max_depth
- **Precision Outcome**: Accuracy not high enough, did not continue to calculating precision

Ensembling

- **Rationale:** Wisdom of the Crowds, improving accuracy by reducing individual model bias, increase robustness by removing outliers and noise,
- **Hyperparameters Tested**: All that were tested during individual model runs
- **Precision Outcome**: Accuracy not high enough, did not continue to calculating precision

Modeling Accuracy and Final Model Selection

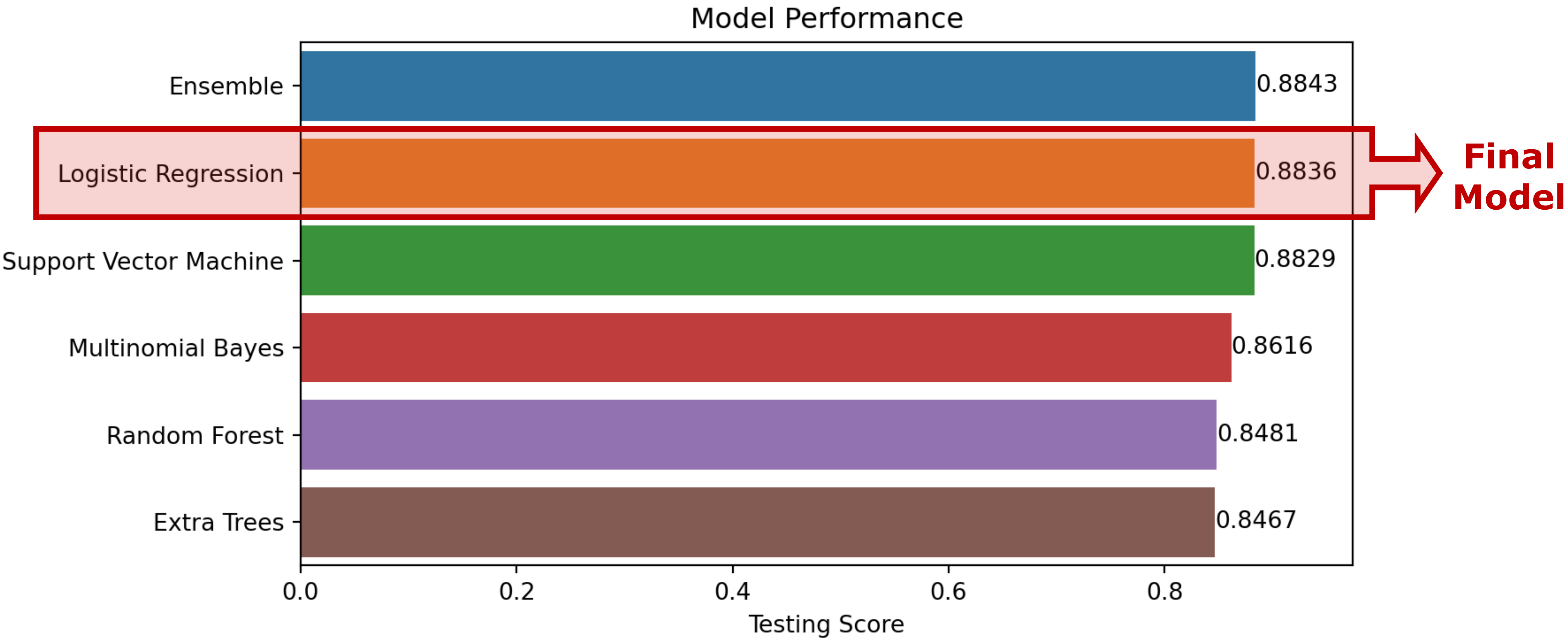
Each model outperformed the baseline significantly, but each model performed similarly in terms of accuracy.



Logistic Regression using TfidfVectorizer transformation was selected to be the final model. While the accuracy was similar to the SVM, we choose the logistic regression for its **simplicity, interpretability, and computational efficiency.**

Modeling Accuracy and Final Model Selection

Each model outperformed the baseline significantly, but each model performed similarly in terms of accuracy.

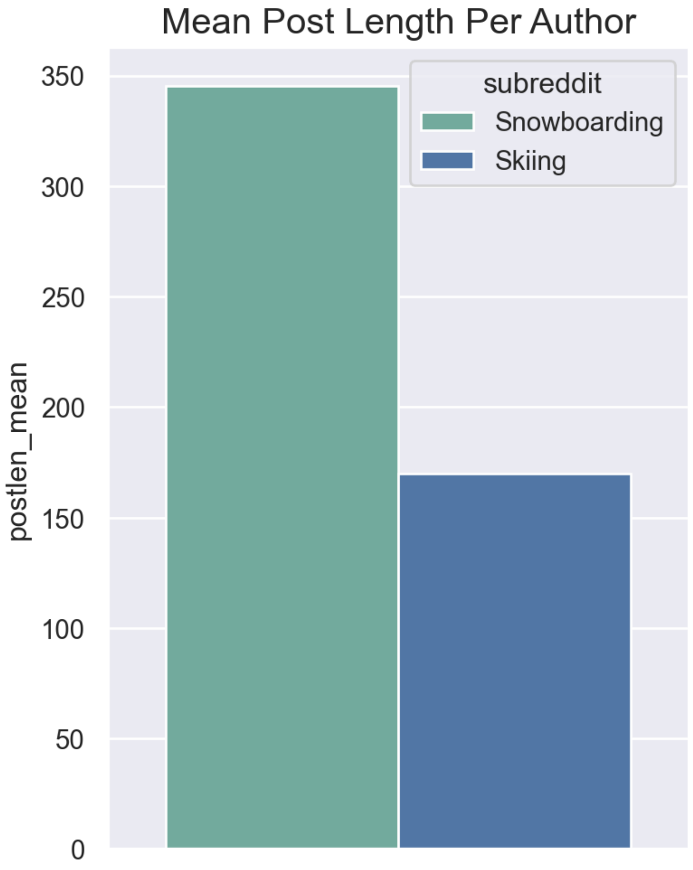


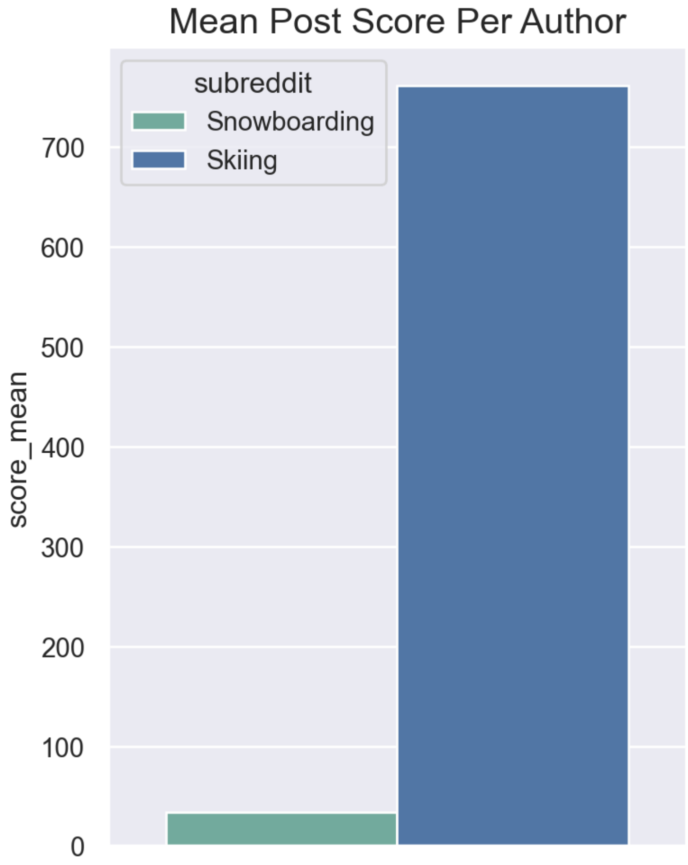
Logistic Regression using TfidfVectorizer transformation was selected to be the final model. While the accuracy was similar to the SVM, we choose the logistic regression for its **simplicity, interpretability, and computational efficiency.**

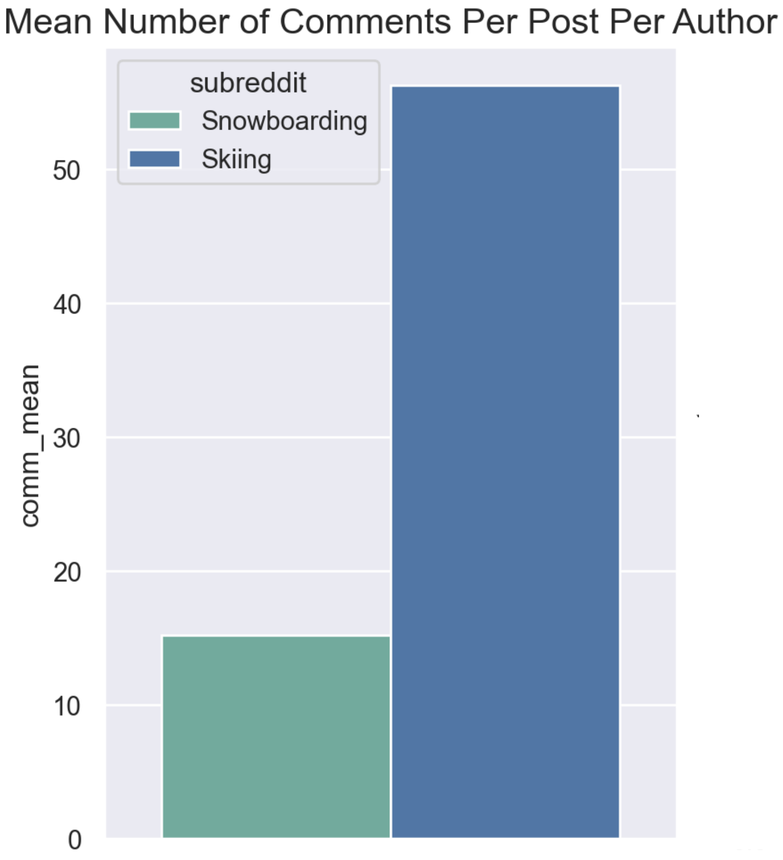
INSIGHT 1

Determining How to Use Our Model: Leverage EDA

To understand how our model can be leveraged to help Burton Snowboards, we must **leverage insights from our exploratory data analysis.**


Mean Post Length Per Author


Mean Post Score Per Author


Mean Number of Comments Per Post Per Author

Both groups had a very similar number of average posts per author. Snowboarders engaged by posting **longer messages** on average and upvoting fewer messages, whereas skiers engaged by **upvoting more** messages and **commenting more** with each post.

INSIGHT 2

Determining How to Use Our Model: Leverage EDA (Cont'd)

Community Engagement

Snowboarders

Top words among the snowboarding community had more to do with gear, performance, technique, or the experience ***in or during*** the activity.

- boot
- Burton
- binding
- edge
- toe
- jump
- heel
- turn
- gear

Skiers

Top words among the skiing community aligned more with the experience ***around*** the activity, such as the best place to go, the best time of year, the best passes to get.

- trip
- Palisade Tahoe/ Lake Tahoe
- season
- Jackson Hole, Park City, Winter Park
- Epic pass / Season Pass / Ikon Pass
- best resort
- day
- Vail resort
- mountain

Summary of Recommendations

Using the insights gained on the subreddit communities, we can now put our model to use in the most effective way possible.

Outreach to Snowboarding Members …

Implement digital transformation efforts to increase E-commerce and direct-to-consumer business by doing the following:

Recommendation	Area	Action
1	**Marketing**	Advertise **promotional offers** and **exclusive discounts** to the snowboard community for gear and accessories.
2	**Partnerships**	Partner with professionals and snowboard schools to package **lessons and rentals/equipment** if bought through Burton.
3	**Content Creation**	Create and deliver targeted content (video tutorials, blog articles, social media posts) on **technical tips and tutorials**.

Summary of Recommendations

Using the insights gained on the subreddit communities, we can now put our model to use in the most effective way possible.

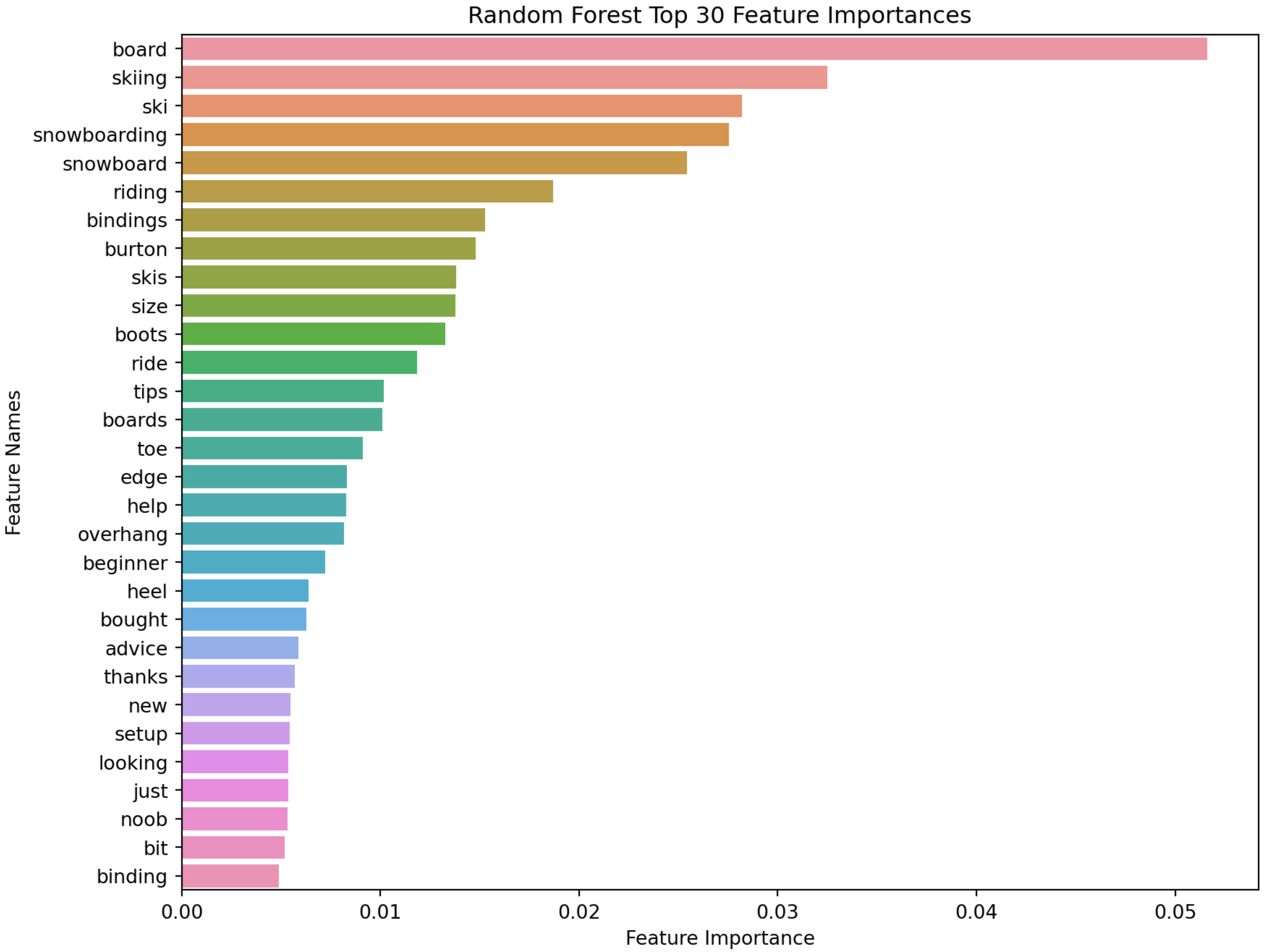
Outreach to Skiing Members …

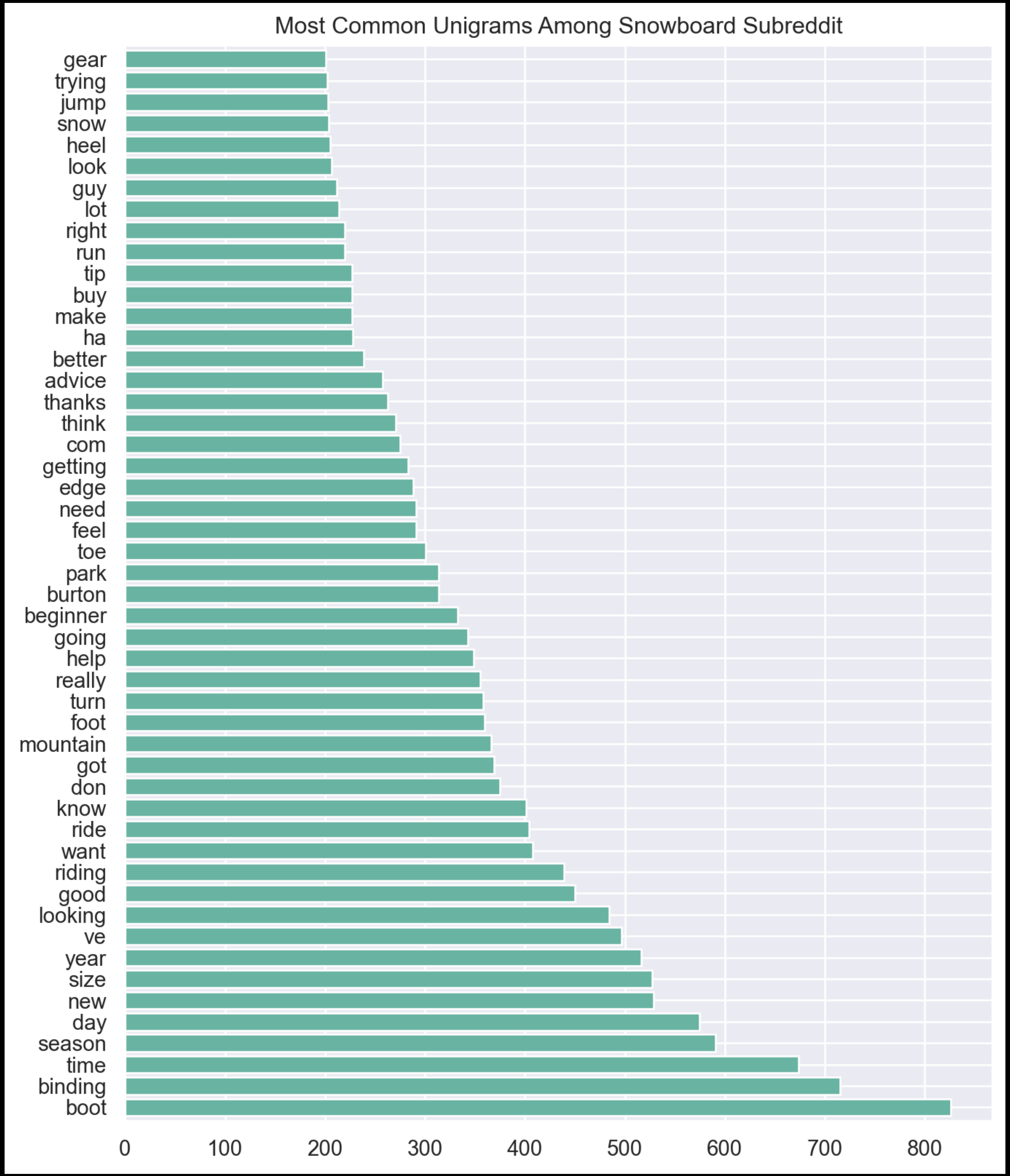
Expand the market share of skiers by advertising Burton’s high-performance apparel and appealing to the community’s interest in trip-planning.

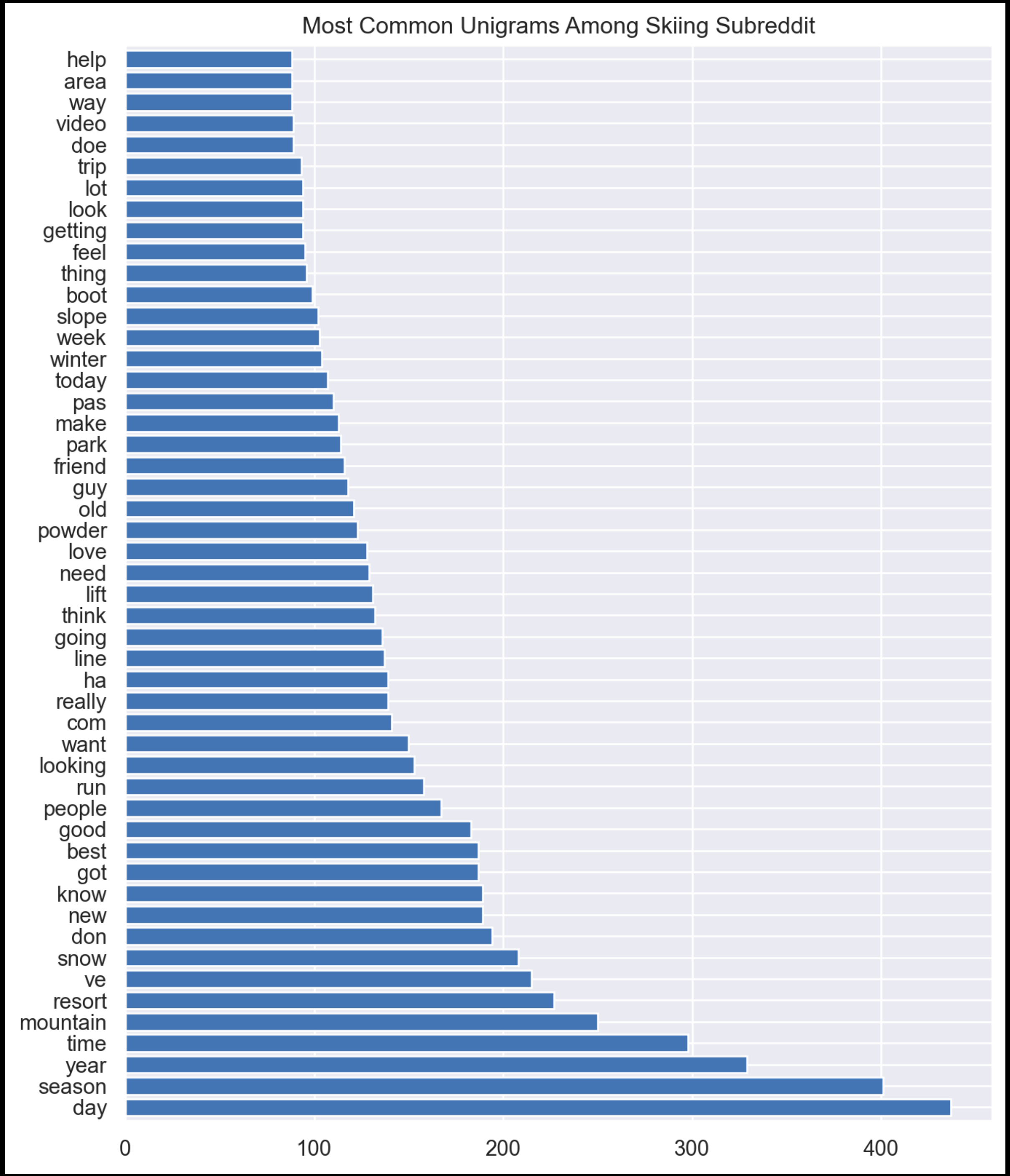
Recommendation	Area	Action
1	**Partnerships**	Partner with **ski resorts and travel agencies** to create exclusive packages that include discounted apparel from Burton. • Use these partnerships to expand Burton’s **global reputation**.
2	**Website Promotion**	Promote skiing section of Burton's website by directing skiers to **special/exclusive apparel and accessory offers**.
3	**Travel / Destination Guides**	Develop guides highlighting the **best ski resorts or locations** for winter activity. Showcase the utility and **advantages of Burton** products as essential gear.

Questions?

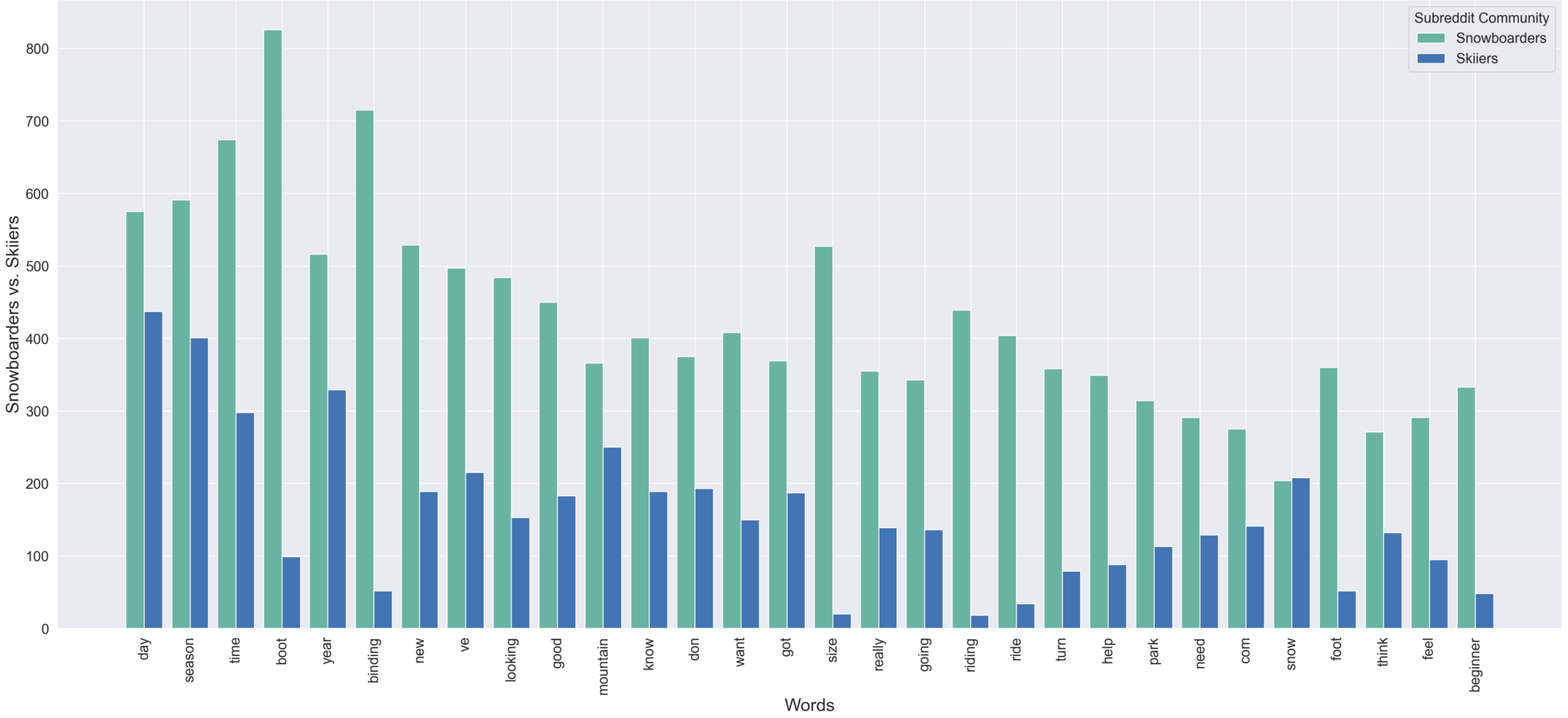
Appendix


Random Forest Top 30 Feature Importances
board
skiing
ski
snowboarding
snowboard
riding
bindings
burton
skis
size
boots
ride
tips
boards
toe
edge
help
overhang
beginner
heel
bought
advice
thanks
new
setup
looking
just
noob
bit
binding
Feature Names
0.00
0.01
0.02
0.03
0.04
0.05
Feature Importance


Most Common Unigrams Among Snowboard Subreddit
gear
trying
jump
snow
heel
look
guy
lot
right
run
tip
buy
make
ha
better
advice
thanks
think
com
getting
edge
need
feel
toe
park
burton
beginner
going
help
really
turn
foot
mountain
got
don
know
ride
want
riding
good
looking
ve
year
size
new
day
season
time
binding
boot
0
100
200
300
400
500
600
700
800


Most Common Unigrams Among Skiing Subreddit
help
area
way
video
doe
trip
lot
look
getting
feel
thing
boot
slope
week
winter
today
pas
make
park
friend
guy
old
powder
love
need
lift
think
going
line
ha
really
com
want
looking
run
people
good
best
got
know
new
don
snow
ve
resort
mountain
time
year
season
day
0
100
200
300
400

Top 30 Words Used in Snowboarding + Skiing Subreddit Communities


Subreddit Community
Snowboarders
Skiers
Snowboarders vs. Skiers
day
season
time
boot
year
binding
new
ve
looking
good
mountain
know
don
want
got
size
really
going
riding
ride
turn
help
park
need
com
snow
foot
think
feel
beginner
Words