Ski Resort Recommendation System Leveraging Data Mining Checkpoint 2

Juncheng Man
University of Colorado Boulder
Boulder, Colorado, USA
juma1543@colorado.edu

Audrey Ly
University of Colorado Boulder
Boulder, Colorado, USA
auly2364@colorado.edu

Grace Waida
University of Colorado Boulder
Boulder, Colorado, USA
grwa6713@colorado.edu

Justin Nguyen
University of Colorado Boulder
Boulder, Colorado, USA
jung1504@colorado.edu

1 INTRODUCTION

Skiing is one of the most popular winter sports in the United States, with a total of 60.4 million skiers reported for the 2023-24 season by The National Ski Areas Association (NSAA) [1]. As the demand for skiing continues to grow, there arises a pertinent problem to find the best ski resort that fits your needs and wants.

Making a choice on the exact resort to visit is difficult as it depends on personal preferences such as location, difficulty, climate, price, and snowfall level. This problem has been compounded by the impacts of climate change, with many ski resorts reduced snowfall and seasons cutting shorter [2].

To address these challenges, we propose a ski resort recommendation system, specifically tailored to help users choose what resort they prefer within the United States. Our system recommends these ski resorts based on user preferences like difficulty, location, price, and travel time, while also taking into account volatile factors like snowfall and weather. Data mining methodologies and tools will be utilized to gain knowledge and extract patterns from past ski resorts and climate data for the assembly of the recommendation system algorithm.

2 RELATED WORK

A number of previous works have focused on developing ski resort recommendation systems, predominantly in the e-commerce sector. An interactive dashboard was developed using data on ski resorts and snow coverage from the Mavens Data playground which was used for the Maven Slope Challenge [5]. The dataset were cleaned and pre-processed using Microsoft Excel, after which it would be stored in a database to be queried by users. The dashboard allows users to narrow down on their ideal ski resort using filter factors like location, season and snow, cost, slope length, and ski lift availability. However, the documentation did not provide evaluation metrics to determine the effectiveness of the solution. Another group developed a deep neural network as a recommendation system that predicts suitable ski resorts for a web agency customers [8]. The deep neural network was trained on a dataset consisting of archived data of Valraiso, a e-Commerce Web Agency of France. It takes in the features: group, budget, duration between order week and stay week, stay week, altitude, ski difficulty, size of skiing area, and country of customer, and outputs a recommended ski resort. The evaluation of the neural network show a performance with a TOP-10 accuracy that reaches 98.95% and 99.30% for French and non-French customers, respectively, where TOP-n accuracy corresponds to the accuracy where the true class matches with any one of the n most probable classes predicted by the model. The NYC Data Science Academy has also published a ski resort recommendation system developed by a student, which helps users determine the optimal ski resort using data scraped from a popular skiing website (www.onthesnow.com) [9]. Two "web spiders" were built via the Scrapy Python framework, with one extracting ski resort data and the other extracting daily snowfall data for each resort. Average annual snow totals, variance in snow totals, cumulative snowfall and lift ticket prices were used to determine the optimal location and timing for skiing. Once again, no evaluation metrics were provided to assess the effectiveness of this solution.

A common pattern across these solutions is the mining of data from resorts and snowfall datasets, and the use of popular features like cost, snowfall, and location to provide recommendations for users. However, none of them has examined in detail the role of climate in shaping the desirability of ski resorts. While the impacts of climate change on snowfall may not have been prominent when the above solutions were developed, in today's world it can no longer be ignored. Gaining knowledge from historical temporal snowfall data is no longer sufficient, and snowfall forecasts have to be combined with climate patterns to provide a more accurate recommendation for ski resorts, as we will do in our project.

3 PROPOSED WORK

3.1 Data Collection

The first step of our analysis is data collection. Our study makes use of open-source datasets from Kaggle, GitHub and government agencies. They include the Ski Resorts and Snow Coverage datasets [7] to understand ski resort types as well as snowfall distribution, Ticket Prices [4] to determine affordability, and Past Weather Data [6], [3] for model construction to predict snowfall for ski locations using live climate data. These datasets are easily downloadable from the respective websites as csv files, which can be further processed using the Python3 programming language in a Jupyter notebook environment.

3.2 Data Aggregation

The second step of our analysis is to aggregate data from the different dataset sources to create a combined dataset where possible for model construction and algorithm design. The aggregation process involves merging data columns from the respective datasets, enabling us to gain insights from a comprehensive view of skiing and provide users with in-depth information related to ski resorts. For model construction, aggregation is performed for the following climate features: temperature, relative humidity, wind speed, wind direction, snow depth. While it is difficult to aggregate the columns of the snowfall and ski resort datasets due to the time-dependence of snowfall dataset and the time-independence of ski resort dataset, we can overlay the datasets visually to analyze them simultaneously.

By integrating our datasets, we aim to deliver high-quality recommendations that consider key factors such as affordability, resort difficulty, snowfall, location, travel time, and weather conditions. Our comprehensive aggregated datasets will be instrumental in uncovering correlations among the various data sources, enabling us to analyze a broad spectrum of information. Leveraging these insights, we can provide users with highly tailored and accurate recommendations based on the identified patterns and relationships.

3.3 Data Pre-processing

The third step of our analysis is data pre-processing, where we ensure the data is of high quality and ready for analysis. This step involves removing duplicates, filling in missing or incomplete information, and cleaning out unnecessary data. This is particularly important when analysing the snowfall and climate datasets, where NaN values or erroneous values appear in bursts likely due to sensor malfunctioning or sensor shutdown. Such data, if untreated, will introduce undesirable bias and noise in our model construction process and must be augmented or removed beforehand. Due to the large size of the snowfall dataset, removal of erroneous and NaN data entries is performed as the cleaned dataset still remains significantly large for model training purposes, totaling up to 23580 entries.

3.4 Knowledge Mining

With the necessary data preparations and cleaning complete, the final step of our analysis is to look at our datasets for valuable patterns and trends to gain insights into how variables pertaining to skiing are related to and influence each other. Understanding the nuanced interrelations in our data guides us to curate the best recommendation algorithm that provides user recommendations based on preferences we believe users would be looking for.

Key patterns to be analyzed include the correlation relationship between features, such as the correlation between cost, skiing difficulty, and location with the popularity of a ski resort, to gain knowledge on the most crucial factors that shape the ski resort choice of the average skier. Such correlation analysis will be performed across all sets of possible features in our dataset to extract the maximum value out of them. In addition, time-series analysis will also be performed to understand the temporal evolution of some features and how relationships between our data features could change over time. We aim to analyze the correlation between snowfall levels and lift ticket prices to determine whether snow conditions influence the cost of skiing at a specific resort. By exploring this relationship, we hope to understand how weather impacts pricing strategies within the ski industry. This can involve exploring the historical trend of snowfall and climate conditions over time, understanding the general seasons where extremities or fluctuations can occur, and using this knowledge to perform time-series forecasting for future predictions.

In addition, data visualization tools like Seaborn and Matplotlib in Python help us visualize the distribution of our data with respect to each other and time. This helps us better identify outliers and trends across complex and multifaceted data like our aggregated dataset. Such visualization may also be provided to the end user for recommendation purposes to help the end user better understand the decision-making process of our algorithm.

3.5 Tools

The tools which we will adopt for the project are as follows.

(1) Jupyter Notebook
(2) Python3
(3) API Services: Google Maps API to estimate the computing distance between the users and the mountain
(4) Microsoft Excel
(5) GitHub
(6) Python Libraries: Pandas, Numpy, Matplotlib, Seaborn, Scipy, Scikit-Learn, xgboost
(7) Kaggle and other datasets

4 EVALUATION

When evaluating users' potential preferences, we consider factors such as ski conditions, weather, ticket prices, route difficulty, and more. From there, we analyze the data to determine what aligns best with user wants and needs. By identifying correlations within the datasets, we have gained valuable insights into patterns and trends. Moving forward, our focus will be on refining these correlations to identify the most relevant and impactful relationships within our data.

(1) **Accuracy of Recommendations:** We will try our best to create insightful graphs that effectively illustrate correlations between key data attributes which will highlight the factors most relevant to users' preferences.
(2) **Cost-Effectiveness:** Our current recommendations provide users with insights into price distributions across ski resorts and the relationship between resort prices and snow coverage. This allows users to make informed decisions based on the correlations identified within the data.
(3) **Potential Mapping Application:** Integrating the Google Maps API is a future goal with the idea of considering traffic conditions and travel times using the most up to date information. For now, we will be using supplemental data for the locations since we are not taking in users' inputs but rather showing the functionality for the time being.
(4) **Ski Run Match:** The system will analyze various types of ski runs (green, blue, black, black diamond, etc.) and assess additional factors such as snowfall, prices, and run lengths. This evaluation will help align recommendations with the user's skill level and specific preferences.

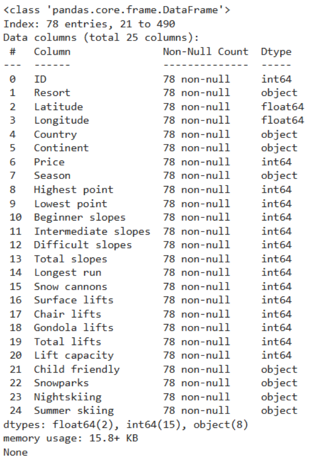
The success of this will be measured on how well the team can interpret the data and provide the users with the best insight possible.

5 MILESTONES

During weeks one and two, we identified a project topic aligned with our interests and refined our datasets to ensure relevance to our problem statement. We also collected the necessary data to address our research questions and prepared a project proposal and presentation. In weeks three through five, we finalized the datasets and gained a deeper understanding of our project's requirements and objectives. This process allowed us to clarify our goals and outline the steps needed to accomplish them. In weeks six and seven, we used Jupyter Notebooks and Python to analyze the data and calculate information from the selected data. This period was focused on starting to use data analysis tools to extract insights from the data. From weeks eight to ten, we conducted a group check-in with our professor to gather feedback and ensure our project was on track. We used this feedback to refine our approach and continued working on data that best supported our goals. Currently, we are finalizing the project and making sure we are projecting the right data based on our insights into the datasets. In the next few weeks, we will begin to prepare our final report and ensure all aspects of the project are met.

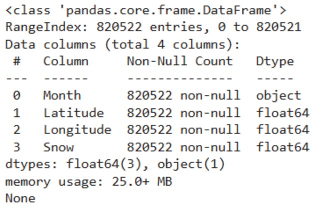
5.1 Completed Milestones

We have collected datasets from Kaggle and various government sources to ensure a robust data foundation. The datasets include the Ski Resorts and Snow Coverage datasets [7] in CSV format and Past Weather Data in both CSV[6] and netCDF[3] format. Data aggregation was performed for the weather datasets to combine the climate variables together with snowfall depth to facilitate model training. Finally, the datasets were cleaned through several pre-processing steps, including removing duplicate and erroneous entries, handling NaN values, discarding irrelevant columns, and filtering relevant rows. Figures 1, 2, 3 shows our datasets after aggregation and cleaning.



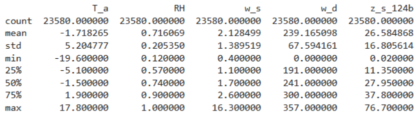
```
<class 'pandas.core.frame.DataFrame'>
Index: 78 entries, 21 to 490
Data columns (total 25 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ID                   78 non-null     int64  
 1   Resort               78 non-null     object 
 2   Latitude             78 non-null     float64
 3   Longitude            78 non-null     float64
 4   Country              78 non-null     object 
 5   Continent            78 non-null     object 
 6   Price                78 non-null     int64  
 7   Season               78 non-null     object 
 8   Highest point        78 non-null     int64  
 9   Lowest point         78 non-null     int64  
 10  Beginner slopes      78 non-null     int64  
 11  Intermediate slopes  78 non-null     int64  
 12  Difficult slopes     78 non-null     int64  
 13  Total slopes         78 non-null     int64  
 14  Longest run          78 non-null     int64  
 15  Snow cannons         78 non-null     int64  
 16  Surface lifts        78 non-null     int64  
 17  Chair lifts          78 non-null     int64  
 18  Gondola lifts        78 non-null     int64  
 19  Total lifts          78 non-null     int64  
 20  Lift capacity        78 non-null     int64  
 21  Child friendly       78 non-null     object 
 22  Snowparks            78 non-null     object 
 23  Nightskiing          78 non-null     object 
 24  Summer skiing        78 non-null     object 
dtypes: float64(2), int64(15), object(8)
memory usage: 15.8+ KB
None
```

Figure 1: Filtered US Ski Resorts Dataset Summary



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 820522 entries, 0 to 820521
Data columns (total 4 columns):
 #   Column     Non-Null Count   Dtype  
---  ------     --------------   -----  
 0   Month      820522 non-null  object 
 1   Latitude   820522 non-null  float64
 2   Longitude  820522 non-null  float64
 3   Snow       820522 non-null  float64
dtypes: float64(3), object(1)
memory usage: 25.0+ MB
None
```

Figure 2: Snowfall Distribution Dataset Summary



```
               T_a            RH           w_s           w_d      z_s_124b
count  23580.000000  23580.000000  23580.000000  23580.000000  23580.000000
mean      -1.718265      0.716069      2.128499    239.165098     26.584868
std        5.204777      0.205350      1.389519     67.594161     16.805614
min      -19.600000      0.120000      0.400000      0.000000      0.020000
25%       -5.100000      0.570000      1.100000    191.000000     11.350000
50%       -1.500000      0.740000      1.700000    241.000000     27.950000
75%        1.900000      0.900000      2.600000    300.000000     37.800000
max       17.800000      1.000000     16.300000    357.000000     76.700000
```

Figure 3: Cleaned and Aggregated Weather and Snowfall Dataset Summary

For knowledge gaining, we first explored the features and their interrelations within each dataset to understand the distributions of the features better. For instance, since the price is likely to be a significant factor in recommending a suitable ski resort, we looked at the distribution of ticket pricing across all of the ski resorts in the United States, highlighted in figure 4.

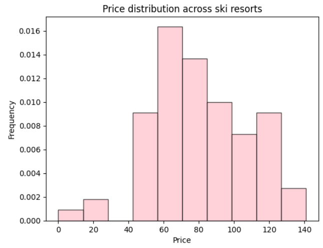


Figure 4: Price Distribution across Ski Resorts

Here, we see that most ticket pricing is clustered about the $50 - $100 range, with very few ski resorts falling below $40. This knowledge helps us to understand that $50 - $100 is what majority of skiers are expecting for ticket pricing when they decide to go skiing, and also that prices in the range below $40 are rare to come by, and definitely should be given higher emphasis by our recommendation algorithm for skiers who value affordable skiing destinations.

We wanted to see the distribution of pricing not only numerically but also geographically. Hence, we also investigated how geographical location and region across the United States may affect the pricing of ski resorts. Figure 5 shows a map of the United States with the various ski resort locations labeled and color-coded by pricing.

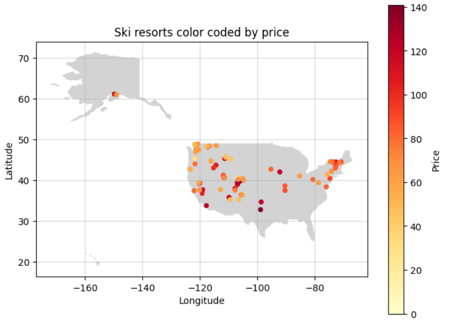


Figure 5: Geographical Distribution of Ski Resort Prices

This exploration reveals that some of the priciest ski resorts are located along the main mountain ranges of the United States, with good slopes that provide a suitable terrain for skiing. This knowledge indicates that the physical terrain of the ski resort is most certainly a strong contributor to the ticket pricing and must be accounted for by our recommendation algorithm. This is further backed up by our exploration of Figure 6, which reveals that some of the priciest resorts do occupy areas with the highest points for skiing. Hence, the recommendation algorithm must reflect that a user who looks for the best slopes and highest points for skiing cannot expect too cheap ticket pricing.

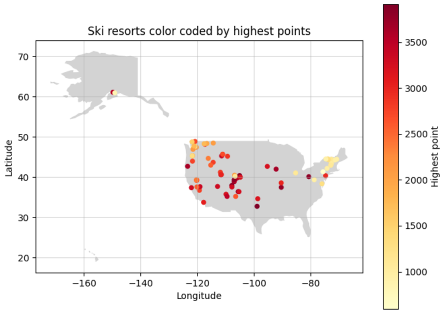


Figure 6: Geographical Distribution of Ski Resort Highest Points

We further explored how ticket pricing could be affected by other features, such as the services and amenities offered by the ski resort itself. To do this, we constructed a correlation heatmap showing the level of correlation between ticket pricing and different features of the ski resort, as shown in Figure 7.

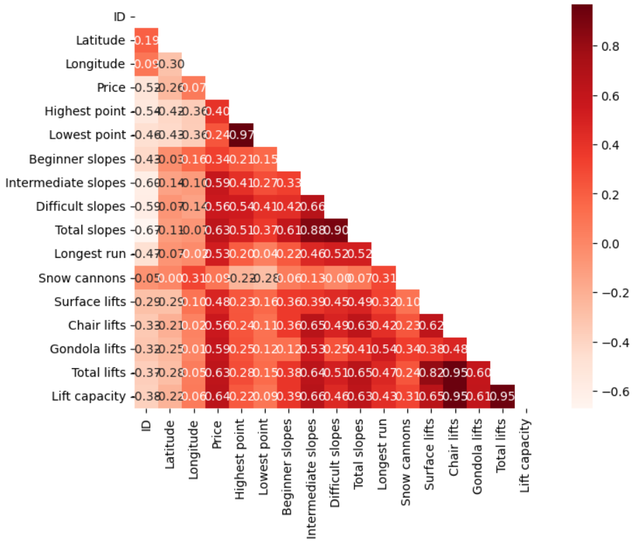


Figure 7: Correlation Heatmap of Ski Resort Features

The heatmap reveals that ski resort pricing is most strongly correlated with both the lift capacities of the resort and the number of slopes, suggesting that the quality and quantity of facilities offered by the ski resort also influence the pricing of the resort on a level similar to that of the natural terrain of the ski resort. This will be factored into the recommendation algorithm for when skiers emphasize the availability of ample facilities in the resort.

Moving on, we also considered how to categorize the many ski resorts in the United States into a few main categories, which could help skiers to zone down on a particular category they may favor quickly, and greatly reduce the search space needed to find an optimal resort by narrowing the search to only that category. This can not only save time for skiers but also resources for the calculations. We approached this through a K-means clustering method to cluster the ski resorts into k clusters, leveraging unsupervised learning techniques. Without knowing the ideal cluster size K, we used the elbow method to first decide on a good value of K, which will be the value of K that gives the turning elbow point in Figure 8.

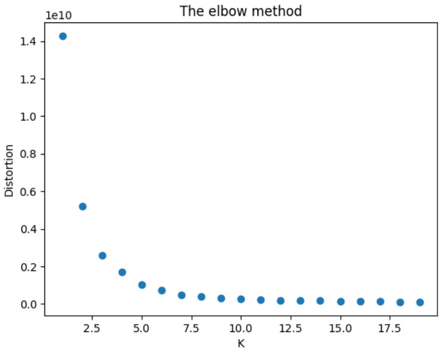


Figure 8: Elbow Method

The graph reveals that the ideal size for K lies around 3 to 4, so we performed K-means clustering using both values of K, shown in Figures 9 and 10.

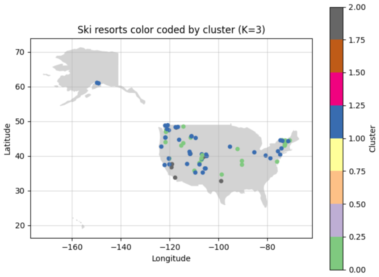


Figure 9: Clustering with K=3

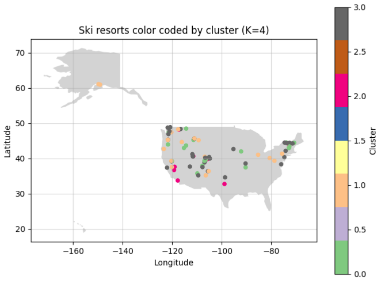


Figure 10: Clustering with K=4

The clustering results show a distinct cluster along the West Coast, with the remaining clusters distributed evenly across the country, suggesting that geography is not a primary factor in clustering. We will investigate deeper to understand the clustering rationales and reveal how the resorts are clustered to help skiers narrow down their search to particular clusters.

We further explored the relation of ski resort pricing with the snowfall they receive, as snow is at the foundation of ski resorts. Thus, intuitively, the pricing should be positively correlated with snowfall, where more snow leads to more skiing activity opportunities. This is investigated by plotting snowfall received across the United States over the period of one year against the ski resorts with their pricing color-coded. The plots are split by month and documented in Table 1. Incorporating the temporal snowfall data reveals an interesting observation: some of the priciest ski resorts do not actually receive substantial snowfall through the year. Some of the ski resorts in the southwest and central regions of the United States have some of the highest prices, but yet the snowfall volume remains on the lower end of the bar. This could imply that such pricing are unwarranted, which must be incorporated into our recommendation algorithm to ensure that such ski resorts are penalized for high prices despite low snowfall levels which translates to less skiing opportunities.

At the same time, we aim to take into account the effects of climate change on snowfall by building a machine learning model capable of predicting based on current weather conditions. Such predictions will be helpful when combined with past snowfall trends to reflect a more accurate projection of snowfall. This is because with the effects of climate change, fluctuations in weather could mean that snowfall patterns deviate from the norm. To construct our model, we made use of the aggregated and cleaned past weather dataset that combines 11 years of snow and weather conditions, making a good spread over time to learn the relationship between weather parameters and snowfall. The model of choice is the xgboost regressor, where squared error is taken as the loss metric. The dataset is then randomly split into a training and testing subset where testing subset takes up 15% of the original dataset. The testing dataset is then further split such that the first 5000 entries are used as validation and the remaining for testing. In order to optimize the hyperparameters for the model, we adopted the GridSearchCV algorithm to iteratively fit and look for the best performing hyperparameters. After 32805 fits, the algorithm result is the following set of hyperparameters: {'alpha': 0.1, 'colsample_bytree': 0.8, 'lambda': 5, 'learning_rate': 0.05, 'max_depth': 6, 'min_child_weight': 5, 'n_estimators': 300, 'subsample': 0.8}. The finally errors for the model are a Mean Absolute Error (MAE)of 12.6818 and Root Mean Squared Error (RMSE) of 15.5756. Figure 11 shows the model's performance on the testing dataset and figure 12 shows the importance of the respective features used by the model, where w_d is wind direction, w_s is wind speed, RH is relative humidity, and T_a is temperature. It is rational that relative humidity has the strongest influence on snowfall since higher humidity indicates more water to form snow. It is however unexpected that temperature has a lower influence.

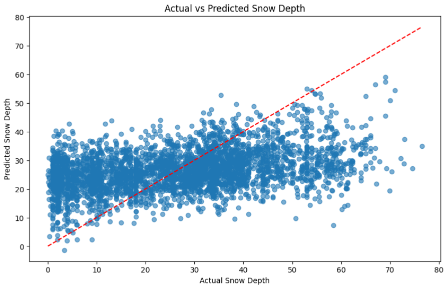


Figure 11: XGBoost Regressor Performance

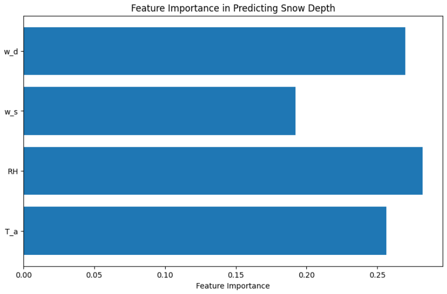


Figure 12: XGBoost Regressor Feature Importance

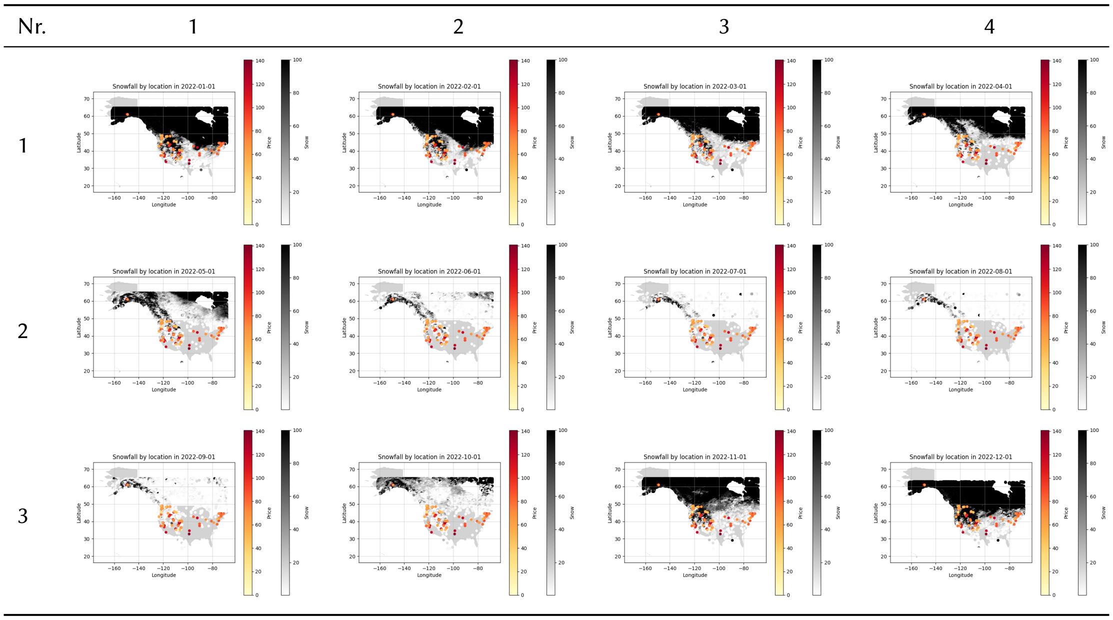

Nr.
1
2
3
4
1
2
3
Snowfall by location in 2022-01-01
Snowfall by location in 2022-02-01
Snowfall by location in 2022-03-01
Snowfall by location in 2022-04-01
Snowfall by location in 2022-05-01
Snowfall by location in 2022-06-01
Snowfall by location in 2022-07-01
Snowfall by location in 2022-08-01
Snowfall by location in 2022-09-01
Snowfall by location in 2022-10-01
Snowfall by location in 2022-11-01
Snowfall by location in 2022-12-01
Latitude
Longitude
Price
Snow

Table 1: Snowfall against Ski Resort Price Distribution (1 year)

The current model still has a relatively high error in the context of snow depth prediction, and figure 11 shows that the model does not perform well at predicting high snowfall depths. We will attempt to address this by using alternate datasets or features to make the model more well-rounded at predictions as an incorrect prediction for low snowfall when there actually will be high snowfall can make a difference in skiing experience.

5.2 Milestones To Do

To enhance our project, we plan to explore the statistics within our current datasets to determine which metrics provide the best insights for users. We'll evaluate additional, more current datasets to include updated information on resort pricing, recent snow depths, and resort distances relative to users. By analyzing how these attributes might impact visitor numbers and return rates, we aim to provide a more user-centric model. For model enhancement, we plan to improve the accuracy of our snow prediction model by incorporating datasets with more influential features. Integrating the Google Maps API will allow users to calculate distances to ski resorts directly, adding convenience.

6 FUTURE IDEAS

Though this is not part of the project scope if we were to move forward with this project, our group envisions expanding this project into a fully developed application that offers users a more interactive and personalized experience. The goal of the app would be to allow users to input specific preferences such as ski run difficulty, budget, and travel distance, and receive tailored ski resort recommendations that align with their needs. If possible, we also want to use and update our datasets, for specifically snow conditions, in real time for our application, to make sure we provide the most up to date information. While we are already working on a system that generates recommendations, this enhanced app would offer more accurate suggestions and a more user-friendly interface, improving the overall experience.

To enhance user experience further, we would integrate additional features such as user reviews, real-time weather updates, hourly snow reports, and possibly even resort-specific services like lift ticket booking. With these improvements, we believe this app has the potential to become the go-to resource for skiing enthusiasts around the world, offering everything from personalized recommendations to live updates and community feedback.

REFERENCES

[1] Forbes. U.s. ski industry had 5th busiest season despite record warm temps in 2023-24, 2024.

[2] Guardian. Ski resorts' era of plentiful snow may be over due to climate crisis, study finds., 2024.

[3] Laboratory, O. R. N. Daymet: Daily surface weather data on a 1-km grid for north america, version 4 r1, 2022.

[4] Lai, S. Onthesnow, 2019.

[5] Malcom, O. C. Find a ski resort for your next vacation: Data analytics project, 2023.

[6] of Agriculture, D. Data from: Eleven years of mountain weather, snow, soil moisture and stream flow data from the rain-snow transition zone, 2019.

[7] Pedersen, U. T. Ski resorts, 2022.

[8] Serdouk, Y., C. T. C. E. . M. C. Ski resorts recommendation using deep neural networks. *Proceedings of the Workshop on Recommenders in Tourism Co-Located with the 15th ACM Conference on Recommender Systems (RecSys 2021) 2974* (2021), 85–89.

[9] Toujas, A. Data web scraping to help plan a ski vacation, 2017.