Justin Rebollo
CS 677
Final Project Summary

Predicting Snowfall based on Historical Data: Snowfall at Hunter Mountain

Introduction

For my project I explore the ability to accurately predict snowfall at ski resorts. While browsing datasets and data science literature, I have not seen many studies on predicting snowfall at ski resorts. The first way to adequately frame the question at hand was to determine the relevant data sources. The dependent variable will be snowfall and the independent variable will be weather related data. After scouring for datasets I was able to find adequate data for snowfall at Hunter Mountain Ski Resort, New York as well as corresponding weather data at OpenWeatherMap and meteomatics.com. I was able to find ten years of historical data. With the appropriate data, I refocused my question further. With ten years of historical data, is it possible to predict out the snowfall for 2021 season at Hunter Mountain? Keep in mind ski season is typically December to April. My final project addresses this question, and ultimately I was able to predict the 2021 snowfall data at a high accuracy using multiple models.

Dataset

The dataset encompasses data from 2010 to 2021 on an hourly basis. The data is comprised of weather features and location data hourly across the dataset. The most important data is the weather data and hourly snowfall. There are roughly 103,000 data points in the set with 25 features.

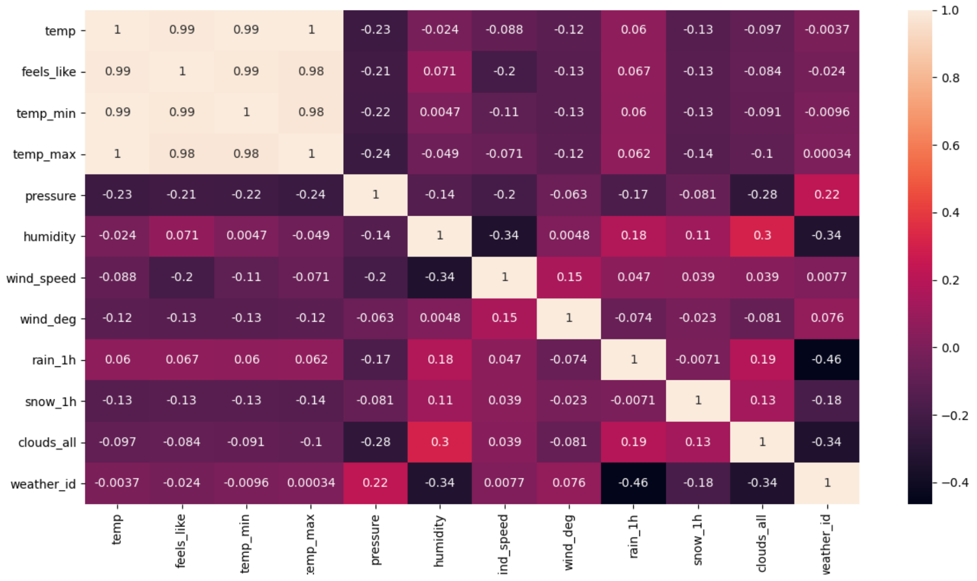
Preprocessing

In order to use this data, I had to preprocess and clean the dataset. Initially, the dataset was full of blanks. Because of this, I filled all of the blanks in my dataset with 0. I label encoded my dependent variable which is hourly snowfall. Then, I scaled the dependent variables. Additionally, in order to split the data to fit the question at hand, I split the data into 2010-2020 and 2021. 2010-2020 is the training set for all models and 2021 is the test set for all models.

Features

The most relevant data to predict the hourly snowfall are: temp_max, humidity, clouds_all, temp_min and wind_speed, and snow_1h. The independent variables for the models are: temp_max, humidity, clouds_all, temp_min and wind_speed. The dependent variable is snow_1h. snow_1h was transformed into a categorical variable to allow for modeling.

Fig 1. Correlation matrix of relevant variables



Modeling, Results

For this project, I implemented four models: Naive Bayesian, Decision Tree, Random Forest and k-nearest neighbors. After implementing each model, I was able to predict the snowfall with a high overall accuracy for the 2021 test data set.

Model	Naive Bayesian	Decision Tree	Random Forest	knn
Accuracy	69.22	88.25	91.08	89.14

The Random forest had the highest overall accuracy of any other model. While knn is the second most accurate model alongside Decision Tree as the third most accurate model. After reviewing the plots a different story is conveyed in terms of the overall accuracy of each model.

Fig 2. Actual and Naive Bayesian predicted snowfall values for 2021

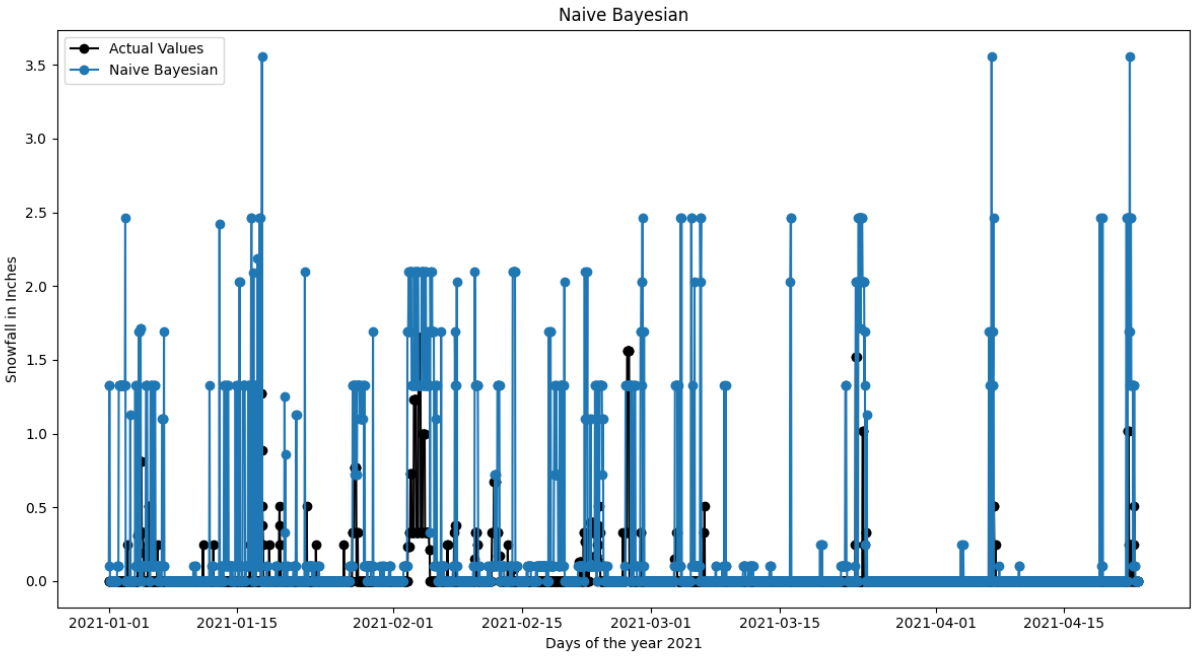


Fig 3. Actual and Decision Tree predicted snowfall values for 2021

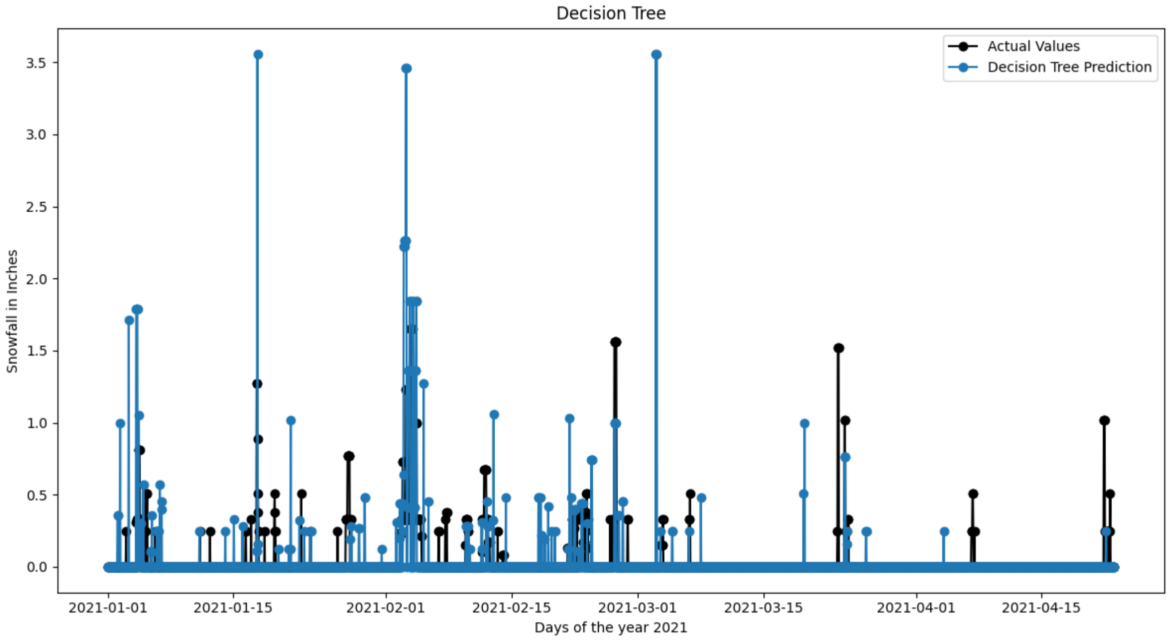


Fig 4. Actual and Random Forest predicted snowfall values for 2021

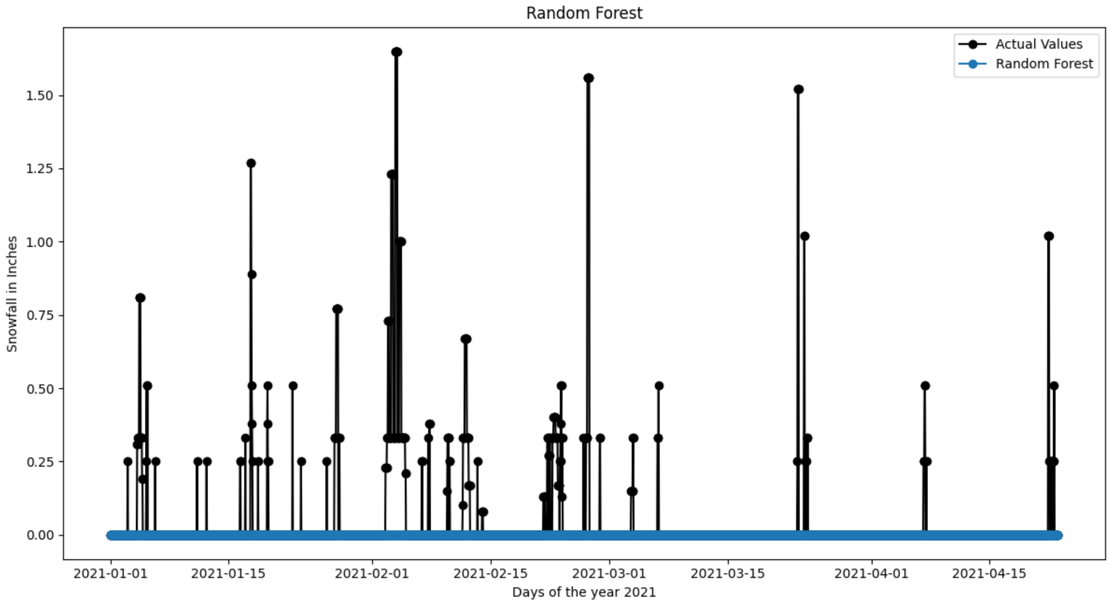
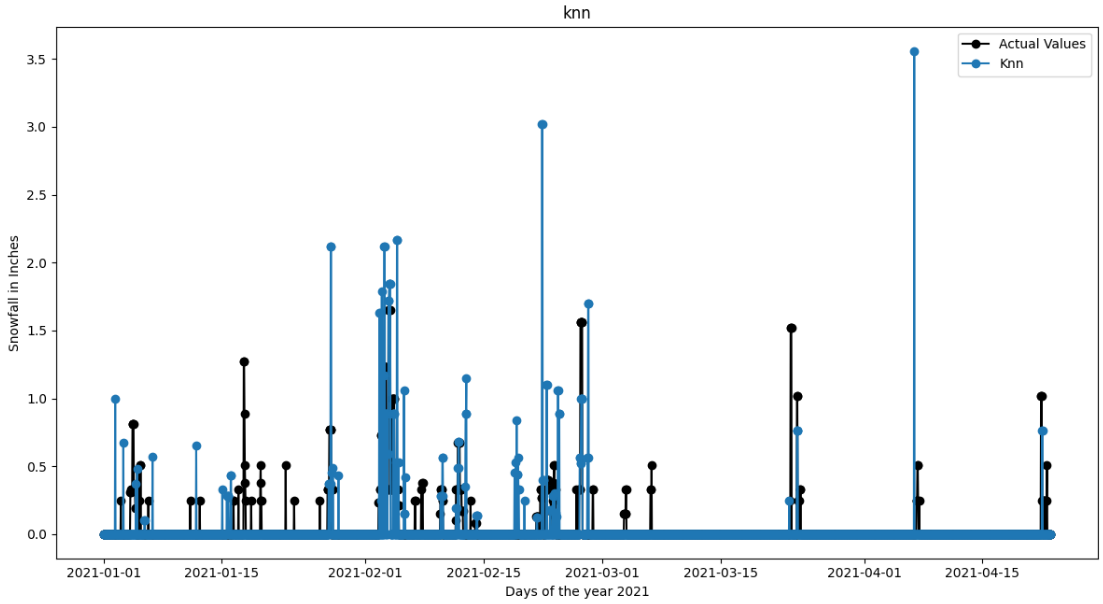


Fig 5. Actual and knn predicted snowfall values for 2021



Discussion

The plots above contain the inverse transformed data from the model's predictions in order to have snowfall in inches as the dependent variable. The snowfall hourly predicted data is initially a categorical value based on range of snowfall values in the training data. The inverse transformed data plotted, reflects the actual snowfall in inches. Overall the models performed relatively accurately. The plots above reflect the actual observed snowfall for 2021 compared to each model's prediction for the snowfall for that day. Fig 2 is Naive Bayesian versus the actual values. It can be seen that the Naive Bayesian model is overfitting the data and is predicting snowfall nearly everyday in the training set at a rate that is nearly. Fig 3 is the Decision Tree model which seems to be accurately reporting the snowfall. Overall, the Decision Tree model predicts at nearly the same frequency as the actual reported values. There is some overproduction of values but largely the Decision Tree model is accurate. By far the most interesting plot from my project is the Random Forest model. The random forest model is 91.08% accurate but if we look at the plot we can see the model predicted no snow for everyday in the targeted interval. This appears to be a mistake until you analyze the data. Overall, snowfall only occurs about 10% of the time in the given time interval. So the Random Forest model is relatively accurate but is not the best model for actually predicting the snowfall. Another highly accurate model is the knn implementation which from the plot, it can be seen that the model preforms in the same manner as the Decision tree.

Conclusions

After analyzing the models and plots I can determine that I was able to create accurate models for predicting snowfall throughout the 2021 ski season. The most interesting result of this project is that the hardest thing to predict is days that is does snow, as the majority of days in the training and test data there is no snow. Using the Decision Tree model and and knn model, we can see that it is possible to predict the snowfall accurately while actually having not all zero values. The overall frequency of snowfall seems to be relatively easy to predict compared to the actual amount of snow. Overall the question at hand was answered with the use of several models. Extending on this in the future, I would focus on reducing the error generation from the high frequency of zero snowfall days. Also, it was a highly successful example of translating continuous variables (snowfall) to a categorical variable in order to predict the amount of snowfall across 2021.