Mohit Bogineni, Gelila Kebede, Aishani Mukherjee, Amman Vahora, and Joseph Yang

CMSC 421 - Introduction to Artificial Intelligence

Professors Sujeong Kim and Max Ehrlich

9 May 2024

Snow Day Prediction Model Report

Introduction and Initial Goals

As a team, we aimed to develop a tool for predicting snow day closures, recognizing the complex factors that influence school closure decisions, such as weather and road conditions. This problem is challenging due to the multitude of variables involved. However, it is valuable as it empowers individuals, particularly parents and children, to better anticipate potential closures at academic institutions, allowing family planning and logistical arrangements. While our original goal was to accurately predict snow day closures, we encountered obstacles related to minimal available data on past closures. We realized that publicly accessible resources such as all school closures across the state of Maryland had conflicts regarding access requests and the deadline for project completion. This required alternative data collection techniques such as scraping through email threads received through one team member and tweets on X manually. We also pivoted our approach to focus solely on one Maryland county. In terms of data collection and tidying, it was based completely on our dataset rather than external resources.

Our project code can be accessed with the following link: HCPSS Closing Predictions

Challenges and Solutions

Our solution to the problem we aim to address is an application designed to predict the probability of a snow day for a given day. We integrated a seamless UI via React and a simple backend via Flask to accompany the model we created to have a friendlier user experience. Initially, we intended to employ Naïve Bayes for its simplicity in classification tasks. However, due to the imbalanced distribution of classes in the problem (only around 3 days a year had a school closure), Naïve Bayes yielded suboptimal results. Specifically, Naïve Bayes would consistently and drasticly over predict the number of snowdays or underpredict the amount of snowdays (the recall would be either close to 0 or close to 1). This resulted in the model not being useful for creating proper informed predictions about a school cancellation on any given day. Instead, we opted for Random Forest, recognizing its simplicity and ability to deliver reasonably accurate outcomes even with poor data quality. It is also better at pinpointing and using only the input variables that were relevant to predicting snowdays.

To develop our solution, we utilized scikit-learn, leveraging its easy-to-implement classification libraries and methods for data splitting and evaluation. scikit-learn's flexibility allowed us to experiment with various models which was crucial for refining our approach. We integrated the OpenMeteo API to access real-time and historical weather data, incorporating factors such as snowfall amount, temperature, daylight time, windchill, windspeed, and precipitation duration to enhance the predictive capability of our model. Our solution combines machine learning techniques with comprehensive weather data to predict snow days accurately while dealing with the limitations posed by imbalanced data distributions and the smaller size of our dataset.

Literature Review and Model Integration

This literature review delves into the development of our snow day prediction prototype, drawing insights from a diverse collection of research. The following research papers examine methodologies, such as imbalance class distributions which are crucial for weather forecasting, that we referred to to create our final product.

The paper "Comparison of Different Machine Learning Models for Short-Term Load Forecasting at Transformer Level with High Amounts of Photovoltaic Generation" by Timon Jungh et al. explores predicting energy loads on transformers that integrate high volumes of solar power. It underscores the importance of including weather data, particularly solar radiation, in predictive models for better accuracy. Utilizing machine learning techniques like CNNs and LSTMs, the study assesses various models' effectiveness over a 24-hour forecast period. Importantly, this study introduced our team to the OpenMeteo API, highlighting its value in prediction models dealing with weather data.

The paper "Classification of Data Streams with Skewed Distribution" by Abhijeet Godase and Vahida Attar explores the challenges of mining data streams with imbalanced class distributions, particularly in applications like fraud detection where minority classes are crucial. The authors propose an ensemble-based method that adjusts the training set with a balanced mix of retained positive examples from previous data chunks. This approach addresses both class imbalance and concept drift, ensuring the model remains effective over time. Their methodology, emphasizing metrics such as AUROC and G-Mean over simple accuracy, could be adapted for developing machine learning models to predict rare events, such as snow days, using skewed weather data.

We also explored a research study by Maria Carolina Monard and Gustavo E.A.P.A. Batista, published in 2003 which also focuses on concept-learning in situations with imbalanced class distributions. The research emphasizes the challenges that such skews present as most traditional models excel at classifying the majority of data while struggling with special cases. This often leads to higher misclassification costs which can be mitigated by utilizing metrics beyond accuracy and error rates, such as false positive/negative rates. Cost-sensitive learning

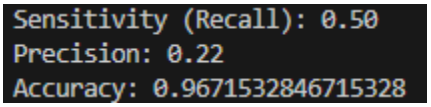
aims to minimize these misclassification costs. Understanding the impact of class distributions on specific learning phases can guide the development of better learning strategies. Initially, we utilized a Naïve Bayes approach to the learning model, however, the results seemed to be suspiciously accurate. However, for specific predictions, the model was greatly overfitting. We discovered that we had been dealing with lop-sided data which the research study by Monard and Batista also mentioned. The research paper mentioned, "there are decision tree splitting criteria that are relatively insensitive to a data set class distributions … similar statements for Naïve Bayes and decision tree learning systems."

Our current model incorporates a similar structure by referring to the research paper "Weather Forecasting Using Machine Learning Algorithm" published at the 2019 International Conference on Signal Processing and Communication (ICSC). The paper introduces a weather forecasting system utilizing real-time data from various features, such as temperature, humidity, etc. The system has a focus on rain prediction whereas our model attempts to examine snow day predictions. The model in the research utilizes Random Forest implemented on a Raspberry Pi in which data is collected and distributed via sensors. The Random Forest model trains to output binary predictions based on the sensor inputs that communicate via GPIO pins (General Purpose Input/Output pins that are utilized on circuits). Our model refers to similar features while using historical data from OpenMeteo to predict snow days instead of sensors and real-time weather.

The paper "Biased Random Forest For Dealing With the Class Imbalance Problem" explores using a modified random forest classifier paired with nearest neighbors to create a biased random forest algorithm which helps solve problems where classifiers tend towards the majority class in cases where the data is heavily imbalanced. The solution explored by the paper tries to move the oversampling from the data level to the algorithm level. When initially training our model, we used a Naïve Bayes algorithm but found that it was overfitting and falling into the problem described by the paper. When we realized this, we switched to using a Random Forest, which yielded better results. The paper describes the modified Random Forest as being fed with more trees from the critical areas. This leads to a more effective way of dealing with the class imbalance problem we had as our data is extremely skewed towards no snow days.

Results and Evaluation

In presenting our results and evaluation, we utilized accuracy, recall, and precision as quantitative metrics to assess the performance of our snow day prediction model. However, accuracy proved somewhat limited in gauging model quality, given the overrepresentation of school days without closings in our dataset, potentially skewing results. For example, with the



dataset containing only around 20 closures for around total days, guessing no closure for every

single day would result in an extremely high accuracy even though the model itself would provide no value. Precision, indicating the proportion of predicted snow days that were confirmed as actual closings, revealed that approximately 22% of predicted snow days aligned with actual occurrences. Meanwhile, recall, showing the percentage of actual snow days correctly identified by the model, indicated that around 50% of all snow day closures were successfully identified. Despite achieving an impressive 97% overall accuracy in determining the presence of closures on any given day, unexpected outcomes and challenges surfaced during the evaluation. Specifically, the scarcity of closures throughout the 6 years from 2014 to 2020 posed difficulties in achieving a fully accurate model. The partitioning of data for the creation of a test set introduced challenges, as the distribution of school closings may not have been evenly dispersed between the training and test sets, potentially impacting evaluation outcomes.

Member Contribution

Overall, our team worked well together in terms of creating the final product and presentation.

For data collection, Gelila, Mohit, and Aishani collaborated on gathering data from various sources and acquiring weather data from API calls via OpenMeteo. Mohit was able to manually gather alerts on snow day closures via email threads and tweets on X to compile data required for the model to train upon.

Mohit and Joseph coordinated the creation of the model and agreed to tweak the model from Naïve Bayes to Random Forest based on results from the evaluation stage. Together, they were able to create an algorithm that has quite a high accuracy by referring to sources mentioned in the literature review section of this report. For this portion, the entire group was consulted for significant changes in the model and how well it was trained.

Gelila and Amman collaborated to create the front-end user interface that would ultimately be integrated with the Random Forest model utilized for this project. They created a beautiful user interface by creating an application through React and Flask. They utilized visualizations that complement the theme of snow in this project and connected the backend to the frontend seamlessly. Group input was used to consult to resolve any issues involving the integration.

As for the presentation, all of us contributed significantly to formatting slides and figures used for the showcase of our application. All five of us have contributed significantly to the creation and appropriate formatting of this project report.

References (APA)

A. Godase and V. Attar, “Classification of data streams with skewed distribution,” *2012 IEEE Conference on Evolving and Adaptive Intelligent Systems*, Madrid, Spain, 2012, pp. 151-156, doi: 10.1109/EAIS.2012.6232821.

M. Bader-El-Den, E. Teitei and T. Perry, “Biased Random Forest For Dealing With the Class Imbalance Problem,” in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 7, pp. 2163-2172, July 2019, doi: 10.1109/TNNLS.2018.2878400.

M. C. Monard, and G. E. A. P. A. Batista, (2002, January). *Learning with skewed class distributions.* Research Gate. https://www.researchgate.net/publication/311396259_Learning_with_skewed_class_distrihutions

N. Singh, S. Chaturvedi and S. Akhter, “Weather Forecasting Using Machine Learning Algorithm,” *2019 International Conference on Signal Processing and Communication (ICSC)*, NOIDA, India, 2019, pp. 171-174, doi: 10.1109/ICSC45622.2019.8938211.

T. Jungh, B. Steinhagen, M. Hesse and K. Schulte, “Comparison of Different Machine Learning Models for Short-Term Load Forecasting at Transformer Level with High Amounts of Photovoltaic Generation,” *2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, Grenoble, France, 2023, pp. 1-5, doi: 10.1109/ISGTEUROPE56780.2023.10407216.