Northeastern University

INFO 7390- ASSIGNMENT 2- FEATURE ENGINEERING WITH SUPERVISED LEARNING

EMMANUEL CHIBUA
College of Engineering
002799484 | chibua.e@northeastern.edu

Table of Contents

Executive Summary

This project aimed to analyze and predict extreme weather events in Ontario, Canada, using historical weather data collected from various observation stations in Ontario since the late 1800s. The data, which included a variety of meteorological measurements such as temperature, precipitation, snowfall, and sunshine hours, was used to engineer a target variable, 'ExtremeWeather', based on certain conditions in the dataset.

The project involved several stages including data collection, cleaning, exploratory data analysis (EDA), and feature engineering. These stages ensured the data was relevant, accurate, and prepared for the subsequent machine learning process.

The core of the project was the development of a machine learning pipeline for predicting extreme weather events. This pipeline included data preprocessing, model training, hyperparameter tuning, and model evaluation. Three different models - Random Forest, Support Vector Machine (SVM), and Logistic Regression - were trained and evaluated.

The performance of each model was assessed using various metrics and visualizations, including classification reports, ROC curves, learning curves, confusion matrices, precision-recall curves, and feature importance (for the Random Forest model). These evaluations provided a comprehensive understanding of each model's strengths, weaknesses, and predictive power.

In conclusion, this project not only developed a robust machine learning model for Extreme Weather Event Detection but also provided valuable insights into the weather patterns in Ontario, Canada. The methodologies and findings from this project are documented thoroughly, providing a solid foundation for future work in this area.

Understanding and Definition of the Problem

The problem is to predict extreme weather events based on a combination of historical weather data and various weather parameters. The objective is to develop a predictive model that can accurately classify whether a day had extreme weather or will experience extreme weather based on these parameters. The target variable for this problem is 'ExtremeWeather', which is a binary indicator denoting whether or not an extreme weather event occurred or is expected to occur. This problem requires a careful understanding of weather patterns and the application of machine learning techniques to make accurate and reliable predictions.

Data Acquisition and Preparation

The data was collected from multiple CSV files scrapped from https://dd.weather.gc.ca/ , each containing weather data from a specific observation station. The files were compiled into a single DataFrame for ease of analysis. The data included various features such as station name, longitude, latitude, climate ID, province, year, month, and several meteorological measurements.

Data Cleaning

The cleaning process involved handling missing values and outliers. Missing values were imputed using appropriate strategies such as forward fill for time series data and zero fill for count variables. Outliers were handled based on the nature of the data and the specific requirements of the model.

Dataset Description

Station Name	Longitude	Latitude	Climate ID	Province	Year	Month	Tm	DwTm	D	Tx	DwTx	Tn	DwTn	S	DwS	S%N	P	DwP	P%N	S_G	Pd	BS	DwBS	BS%	HDD	CDD
EMO	-93.800	48.633	6022300	ON	1944	1	-8.7	0	NA	8.9	0	-33.3	0	0.0	0	NA	8.4	0	NA		1			NA	828.0	0.0
EMO	-93.800	48.633	6022300	ON	1944	2	-14.1	0	NA	0.6	0	-40.0	0	4.8	0	NA	4.8	0	NA		3			NA	929.7	0.0
EMO	-93.800	48.633	6022300	ON	1944	3	-8.5	0	NA	5.0	0	-36.7	0	52.2	0	NA	52.2	0	NA		10			NA	821.4	0.0

Figure 1- Snippet of the dataset

The dataset used in this project (with 27 columns and over 160,000 rows) is a collection of historical weather data, which includes a variety of meteorological measurements. These measurements encompass features such as Mean Temperature (Tm), Maximum Temperature (Tx), Minimum Temperature (Tn), Precipitation (P), Snowfall (S), Heating Degree Days (HDD), and Cooling Degree Days (CDD).

The target variable for this project, ExtremeWeather, is a product of feature engineering. It is derived from the existing features in the dataset and is designed to indicate whether an extreme weather event occurred on a given day. This binary classification forms the basis of the problem we are trying to solve - predicting extreme weather events.

The dataset is divided into two subsets for the purpose of model training and evaluation. 80% of the data is used for training the model, and the remaining 20% is reserved for testing the model's performance.

To ensure that the models are not unduly influenced by the scale of the features, the features are standardized using the StandardScaler from sklearn. This process involves removing the mean and scaling the features to unit variance, which allows the models to learn from the underlying patterns in the data more effectively.

Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution of the data, identify patterns, and detect relationships among variables. Various visualization techniques were used including histograms, box plots, and correlation matrix heatmap. The EDA revealed interesting insights such as seasonal patterns in temperature and precipitation.

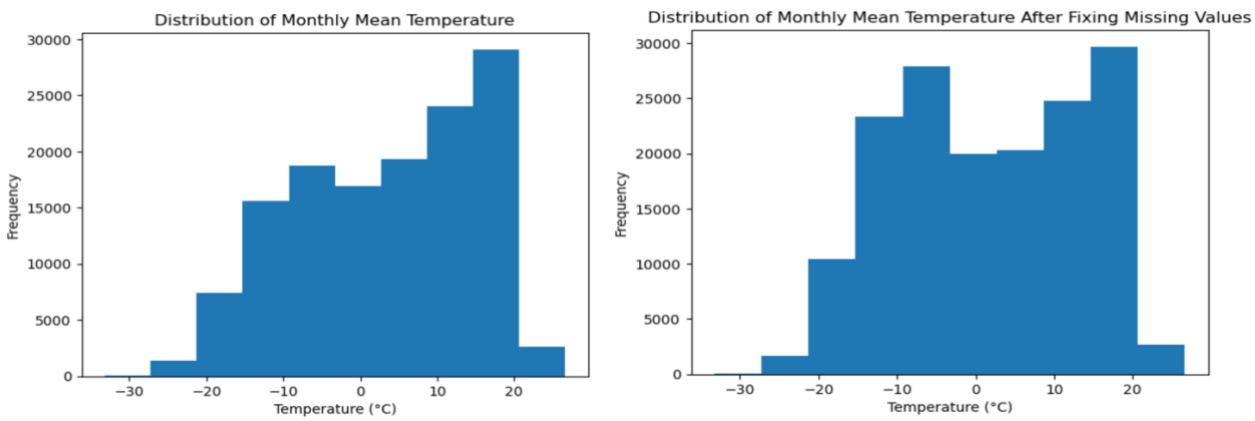


Figure 2- Histogram plots for Mean Temperature (Tm) before and after fixing missing values.

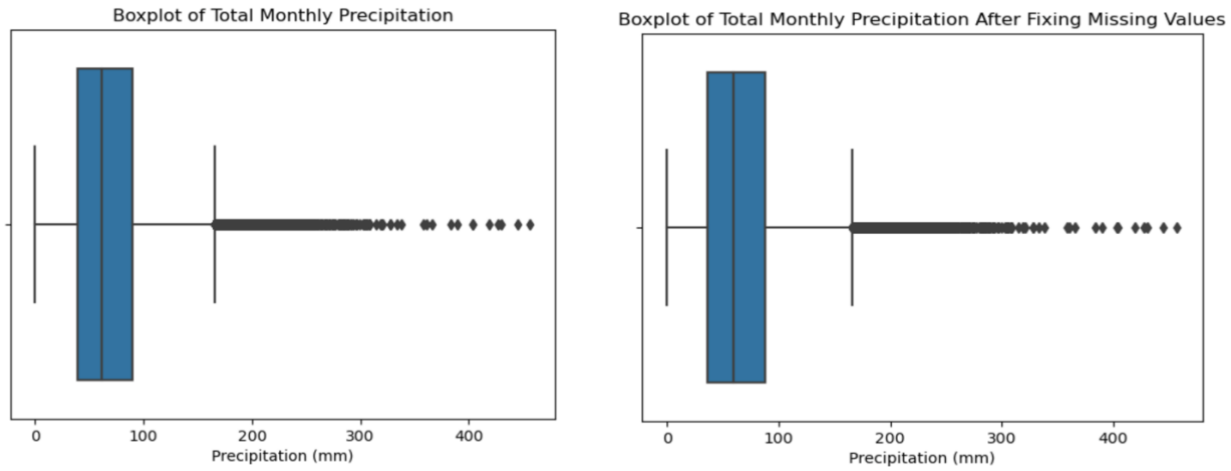


Figure 3- Boxplots of Total Monthly Precipitation (P) before and after fixing missing values.

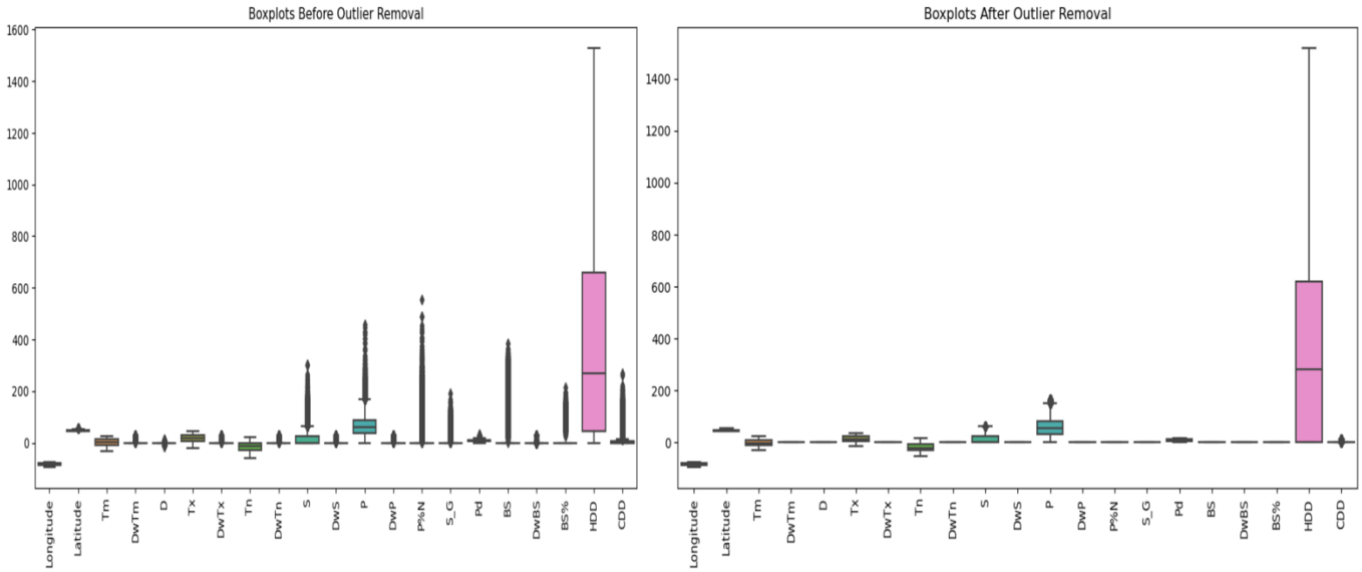


Figure 4- Boxplots of Outliers detection before and after removal

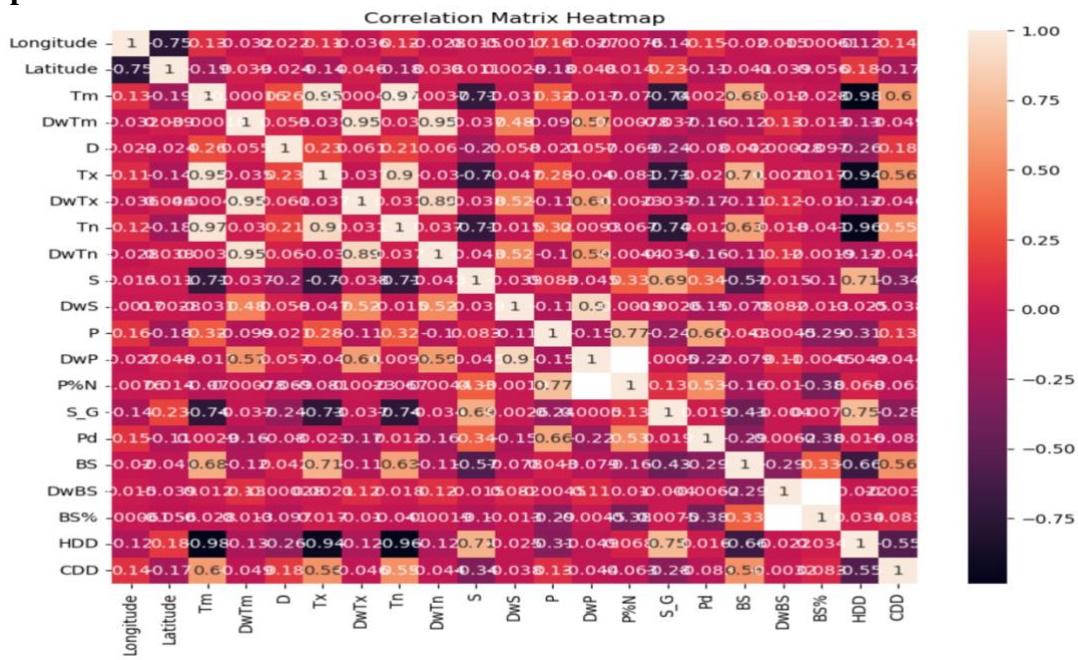


Figure 5- Correlation Matrix heatmap of features

Feature Engineering

In this project, feature engineering played a crucial role in enhancing the predictive power of the model. The process involved the creation of a target variable, selection of relevant features, and transformation of variables.

Target Variable: The target variable, ExtremeWeather, was engineered based on the Tn (Minimum Temperature), Tm (Mean Temperature), and P (Total Precipitation) columns in the dataset. In addition- **Season**, **Climate_Zone**, **Temp_Range**, and **Tm_Binned** were also engineered.

Feature Selection: Features were carefully selected based on their relevance to the problem at hand. This ensured that the model was provided with meaningful input for making accurate predictions.

Handling Categorical Variables: Categorical variables were transformed into a format that could be understood by the model using encoding techniques. It's important to note that all categorical variables were assumed to be already encoded.

Scaling and Normalizing Numerical Features: Numerical features were scaled and normalized using the StandardScaler from sklearn. This step is crucial in ensuring that all numerical features have the same scale, preventing any one feature from dominating others due to its scale.

Through these steps, the feature engineering process aimed to create a dataset that could effectively train the model to predict extreme weather conditions. The process ensured that the model had access to relevant, meaningful, and appropriately scaled features.

Model Selection and Development

Three supervised learning algorithms were selected for this problem: Random Forest Classifier, Support Vector Machines (SVM), and Logistic Regression.

Each model was trained using the training set, and hyperparameters were tuned to optimize the model's performance. For each model, a GridSearchCV is performed to find the best hyperparameters from a predefined set. The best model is then retrained on the training data.

Model Evaluation

Each model's performance was evaluated using appropriate metrics, including accuracy, precision, recall, and F1-score. The models were validated using the testing set, and their strengths and weaknesses were analyzed.

The Receiver Operating Characteristic (ROC) curve for each model is plotted, which is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

A learning curve is generated for each model, which shows the validation and training score of an estimator for varying numbers of training samples.

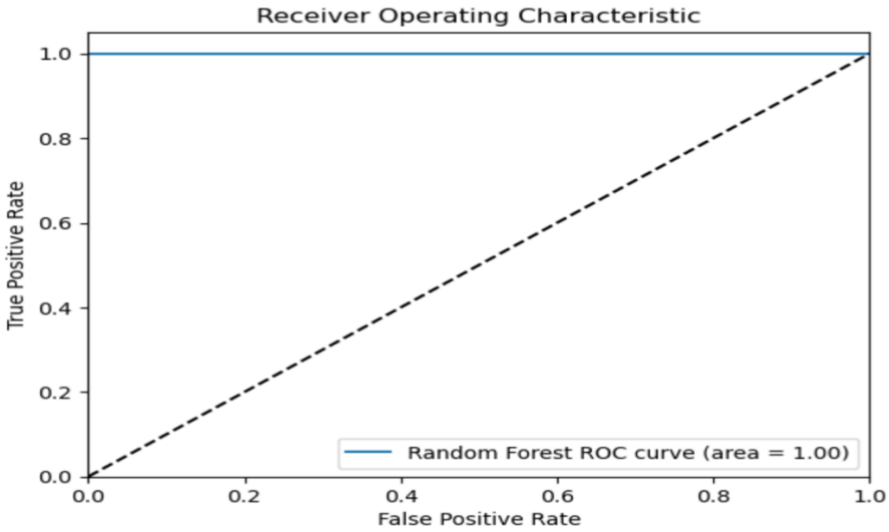
A confusion matrix is generated for each model, which provides a more detailed breakdown of correct and incorrect classifications for each class.

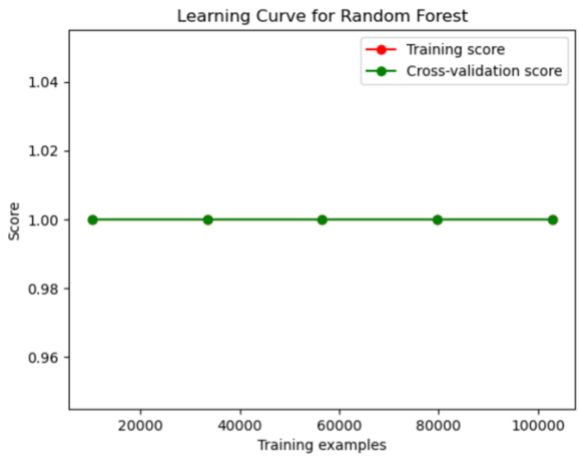
A precision-recall curve is also plotted for each model, which shows the tradeoff between precision and recall for different threshold settings. Below images elaborate more the model performance and evaluation for each of the three supervised learning algorithms used in this project.

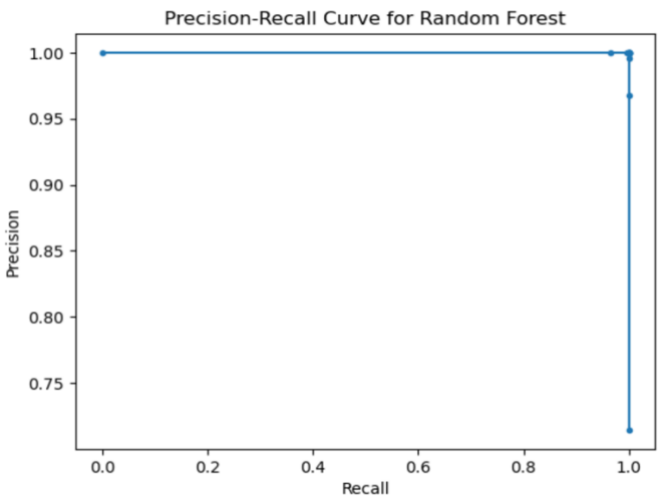
Random Forest Classifier

- **Strengths:** The model's perfect scores in all evaluation metrics suggest that it is highly effective for the given dataset. The model's best parameters, including a maximum depth of 'None' and minimum samples split of '2', suggest that it is well-tuned.
- **Weaknesses:** While the model's performance on the given test set is impressive, there is a risk of overfitting since all evaluation metrics are perfect. It would be crucial to test this model on more diverse datasets or apply cross-validation techniques to confirm its robustness. Additionally, it's essential to consider feature importance and potential biases in data collection or labeling processes to ensure comprehensive evaluation.

```
Random Forest Best Parameters: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 50}
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      9177
           1       1.00      1.00      1.00     22970

    accuracy                           1.00     32147
   macro avg       1.00      1.00      1.00     32147
weighted avg       1.00      1.00      1.00     32147

Random Forest ROC AUC Score: 1.0
```








```
Random Forest Confusion Matrix:
[[ 9177     0]
 [    0 22970]]
```

Figure 6- Random Forest Classifier Evaluation metrics (ROC, Precision-Recall and learning curve)

Support Vector Machines (SVM)

- **Strengths:** The model's near-perfect scores in all evaluation metrics suggest that it is highly effective for the given dataset. The model's best parameters, including a C value of 10 and a gamma value of 1, suggest that it is well-tuned.
- **Weaknesses:** While the model's performance on the given test set is impressive, there is a risk of overfitting since all evaluation metrics are near perfect. It would be crucial to test this model on more diverse datasets or apply cross-validation techniques to confirm its robustness. Additionally, it's essential to consider feature importance and potential biases in data collection or labeling processes to ensure comprehensive evaluation.

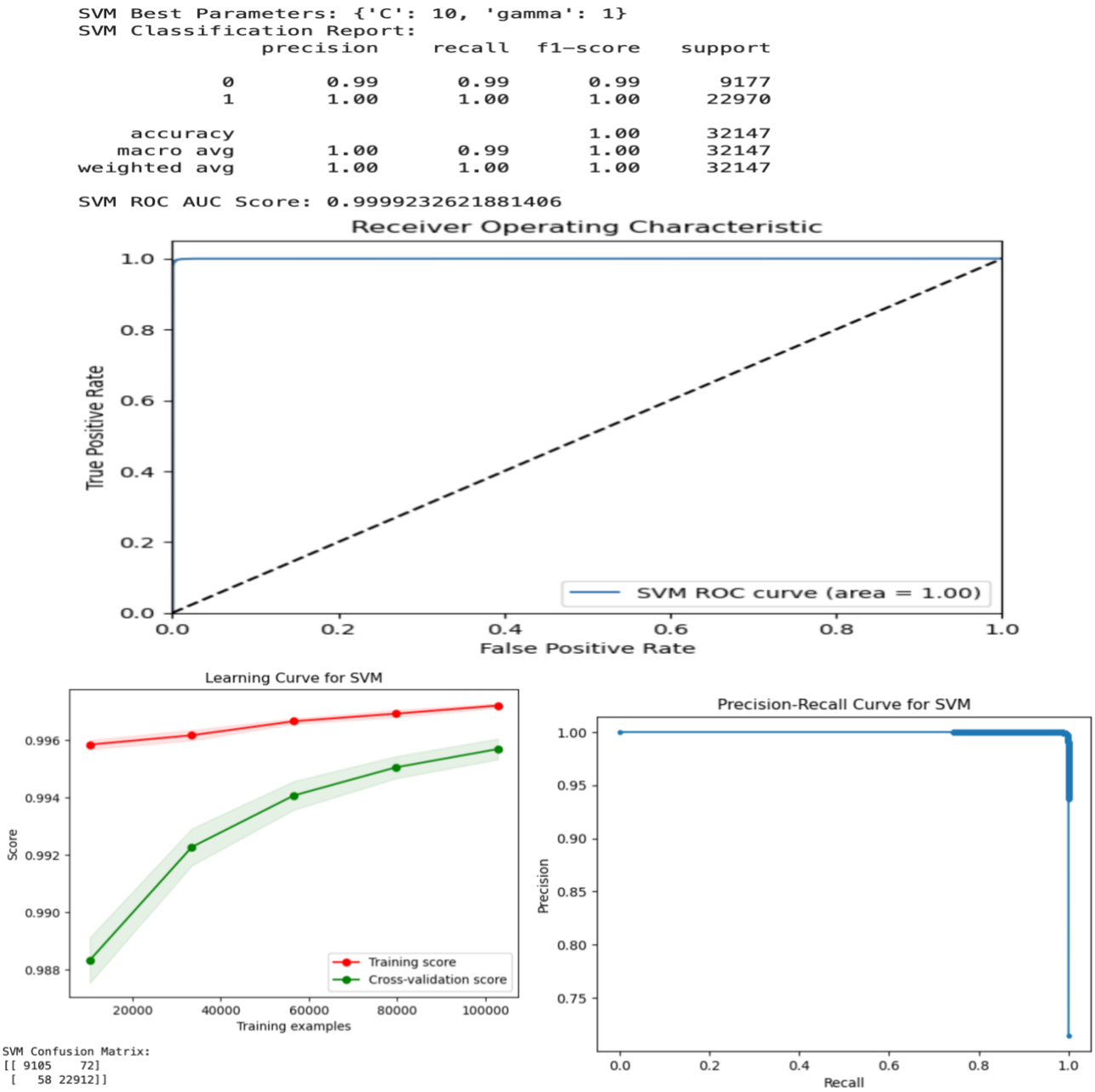


Figure 7- Support Vector Machine (SVM) Evaluation metrics (ROC, Precision-Recall and learning curve)

Logistic Regression

- **Strengths:** The model's high scores in all evaluation metrics suggest that it is effective for the given dataset. The model's best parameters, including a C value of 10, suggest that it is well-tuned.
- **Weaknesses:** While the model's performance on the given test set is impressive, there is a risk of overfitting since all evaluation metrics are high. It would be crucial to test this model on more diverse datasets or apply cross-validation techniques to confirm its robustness. Additionally, it's essential to consider feature importance and potential biases in data collection or labeling processes to ensure comprehensive evaluation.

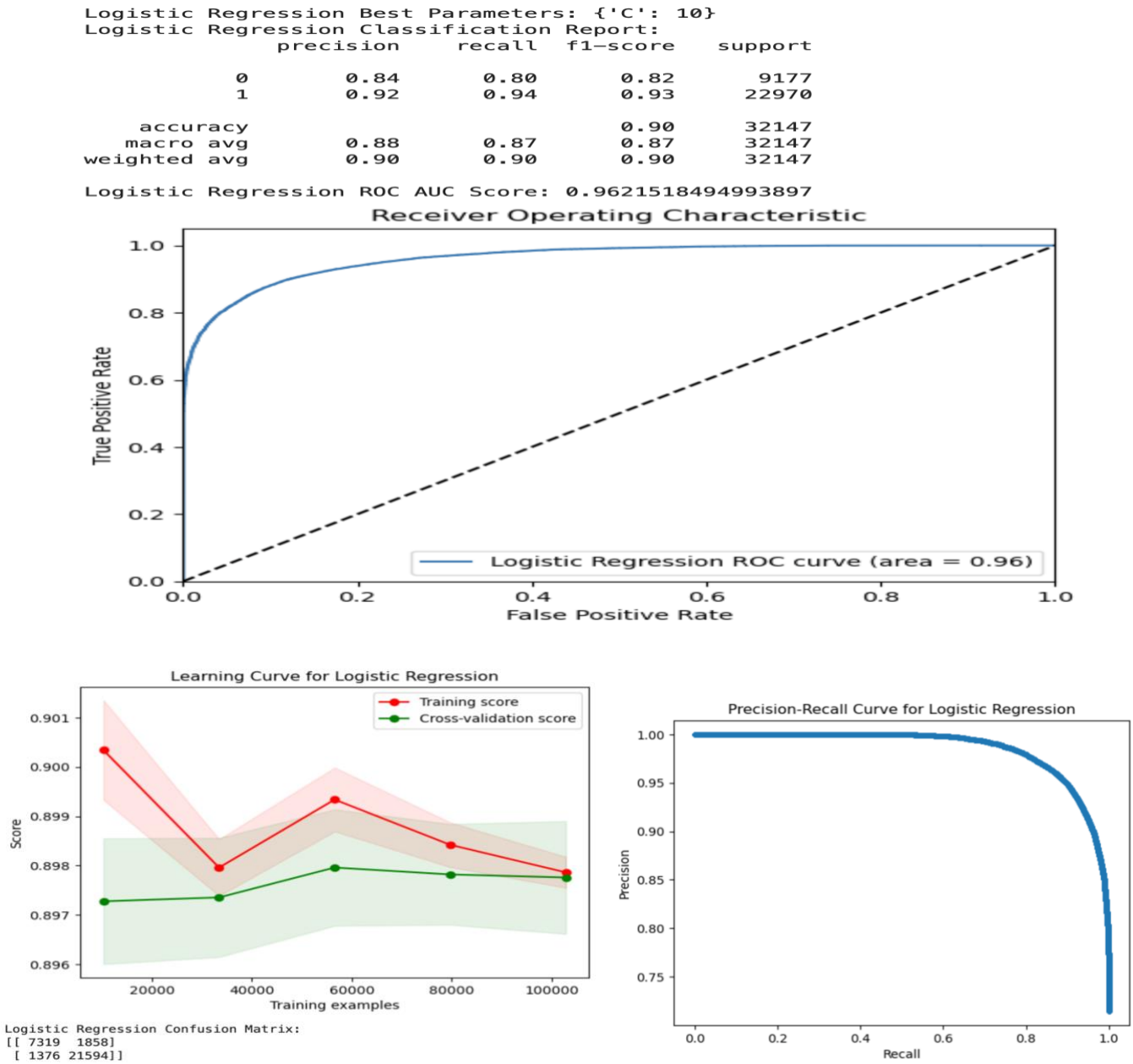


Figure 8- Logistic Regression Evaluation metrics (ROC, Precision-Recall and learning curve)

Ethical Considerations and Fairness

It is necessary to mention that the data used in this data science project belongs to the government of Canada and were ethically accessed from the website https://dd.weather.gc.ca/ . Consideration was given to ethical implications related to the data and the model's application as this project is intended for educational and research purposes. It is important to note that the models should be used responsibly. The predictions made by the models should not be used to make decisions that could negatively impact individuals or communities.

Reflection and Learning

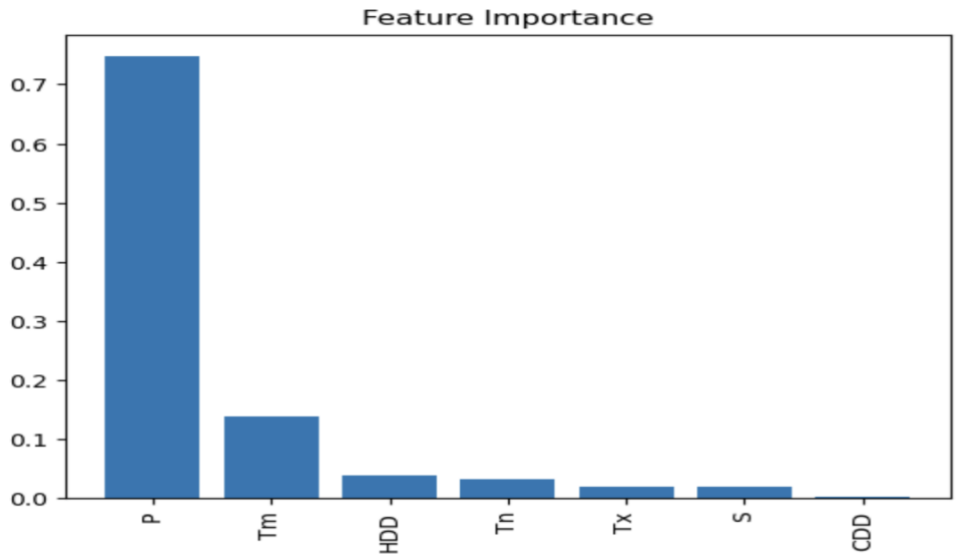
This project provided valuable experience in the end-to-end process of a machine learning project, from understanding the problem and preparing the data to engineering features and developing models. It highlighted the importance of each step in the process and the need for careful consideration and decision-making. This project provided a valuable opportunity to apply and deepen my understanding of supervised learning and feature engineering. I encountered challenges in selecting the appropriate models and tuning their hyperparameters, but I was able to overcome these challenges through research and experimentation. I learned a lot about the importance of data preprocessing and feature selection in improving model performance. I also gained experience in using various metrics and visualizations to evaluate model performance.

Recommendations for Future Work

Future work on this project could involve exploring other models and feature engineering techniques to improve performance. Additionally, more in-depth analysis could be done to understand the relationships between the features and the target variable. Finally, ethical and fairness considerations could be more explicitly addressed in the model development and evaluation process.

Appendices

Feature importance chart for the Random Forest classifier



Data dictionary/ Column data description

- **Station Name**: The name of the weather station.
- **Longitude**: The longitudinal coordinate of the weather station.
- **Latitude**: The latitudinal coordinate of the weather station.
- **Climate ID**: A unique identifier for the climate data from that specific location.
- **Province**: The province in Canada where the weather station is located.
- **Year**: The year when the data was recorded.
- **Month**: The month when the data was recorded.
- **Tm**: Mean temperature for the month.
- **DwTm**: Days with missing mean temperature data in a month.
- **D**: Mean number of days with precipitation >= 0.2 mm in a month
- **Tx**: Maximum temperature for a given month.
- **DwTx**: Days with missing maximum temperature data in a given month.
- **Tn**: Minimum temperature for a given month.
- **DwTn**: Days with missing minimum temperature data in a given month.
- **S**: Snowfall total for a given month (cm)
- **DwS**: Days with missing snowfall total data in a given month.
- **S%N**: Percent of normal (1981-2010) snowfall total received during a given period (%)
- **P**: Total precipitation received during a given period (mm)
- **DwP**: Days with missing total precipitation data during that period
- **P%N**: Percent of normal (1981-2010) total precipitation received during that period (%)
- **S_G**: Snow on ground at end of last day of each period (cm)
- **Pd**: Number days with measurable precipitation (> 0.2 mm) during each period
- **BS**: Bright sunshine hours received during each period.
- **DwBS**: Days without bright sunshine hours information available
- **BS%**: Percent possible bright sunshine hours received (%)
- **HDD**: Heating degree days calculated using 18°C as base value, cumulative over each time interval.
- **CDD**: Cooling degree days calculated using 18°C as base value, cumulative over each time interval.