Jeepal Patel  
November 03, 2021

**Mid-term project report**

# **problem:**

In June 2013, heavy rainfall on the melting snowpack in the Rocky Mountains combined with **Bow River** caused rapid and intense flooding in southern-Alberta watersheds. About 80,000 people were evacuated over the course of the flood. Also, there was $6 billion in financial losses and property damages across the province. Flooding disrupted businesses, damaged critical infrastructures and led to power outages across Calgary. **Forecasting a flood of this magnitude** is a problem that we will attempt to solve here. An accurate flood forecast could have allowed for mitigation procedures to have been initiated, and the extensive damage may have been partially reduced.

To solve this problem, we will need historical data of water flow rates in the Bow River, both in Calgary and upstream in the headwaters of the Bow River. Other data that may come in useful include rain & snow, snowpack depth, river & creek flows, reservoir levels, soil moisture, ware use & demand, and temperatures.

# **methodology:**

Below are the steps we followed to solve the problem and divided into sequential tasks:

1. **Define scope of the project**  
   we as a group decided to focus on predicting flow rates in the Bow River in Calgary with 1 and 2 days of Time Lag. However, we decided to work on 10 years data from 2010 to 2019 inclusive.
2. **Gather Raw Data**from the internet, we were able to find archives of the data for water flow rates in the Bow River, as well as historical weather at weather stations in Canada.

Water flow rates: <https://wateroffice.ec.gc.ca/mainmenu/historical_data_index_e.html>

We identified 17 water flow gauges in Calgary and upstream along the Bow River, that could be important in crafting a solution.   
Weather data:   
<https://climate.weather.gc.ca/historical_data/search_historic_data_e.html>

We identified 2 weather stations, in Banff and Kananaskis, that could be important to predicting water flow rates. Items of interest included air temperature, and precipitation.   
This seemed like enough data to allow us to craft a model of flow rates, and predictions.

**Reformate data**To use the data, we needed to reformat and combine all the data in a single spreadsheet. Placing all the data on a single spreadsheet with each row representing a single date would allow for easier manipulation and integration into mlOS.

**Flow data:** Flow data came as 17 separates .csv files containing data for each individual station for all the dates that the station was in operation (some opened as early as 1909). This data was concatenated and reformatted using custom written MATLAB/Octave s CombineWaterFlowData.m and WriteFlowRatesCSV.m

**Weather data:** Weather data came as individual .csv files containing data for each individual year. This data was manually concatenated in Excel.

At the end, the data was finally manually combined into a single Excel spreadsheet, each row representing a single date with all relevant data in columns.

1. **Create output feature**The feature we were trying to predict (our target) needed to be created. Since we were trying to forecast the flow rate 1 and 2 days in the future, we created targets by selecting the flow rate from the Calgary station 1 and 2 days in the future.
2. **Data wrangling**During the data wrangling steps, first we removed winter months (these months had sporadic water flow rates with some stations frozen and not reporting data/missing values). We kept only May to August, as these are the months with the highest flow rates.

Additionally, we removed missing values from each column, as sometimes the weather stations were not operative.

Lastly, we removed a couple of the river flow stations, as three of them were also missing several years worth of data. (05BA002 - PIPESTONE RIVER NEAR LAKE LOUISE, 05BC001 - SPRAY RIVER AT BANFF, 05BH013 - JUMPINGPOUND CREEK NEAR COX HILL)

1. **Pre-processing data**

We performed min-max normalization on the entire dataset, except for our target, year, month, and day as we wanted to have performance metrics and results that could be readily understood by us.

1. **Split data**

Data was split into a validation and test dataset with at 80/20 split.

1. **Build Machine Learning Model**

A regression model constructed as we were trying to predict a value. Several machine learning models were tested with different regressors used including Random Forest, XGBoost.

1. **Evaluate the Model**

Multiple regressors were tested, with variables adjusted as necessary. The best regressor identified was RndomForestRegressor.

# **results:**

The results of our best model predicted the flood of the Bow River in 2013 in the mlOS. The actual daily flow rate on June 21, 2013, was 1750 m3/s. One day before the maximum flood, our model predicted 1672 m3/s and two days before the flood, 1638 m3/s. This corresponds to errors of 4.5% and 6.4%, respectively.

The features which were most important in the model were BOW RIVER AT CALGARY- 05BH004 (affected by 32%), GHOST TAILRACE- 05BE999 (affected by 30%), and KANANASKIS RIVER BELOW BARRIER DAM- 05BF025 (affected by 8%).

Graphical user interface, chart

Description automatically generated

We were also able to predict the nearest value of the flow as shown in the graph below.

Chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Above graphs also explain the error vs predicted target and predicted vs target. These values are predicted with Time Lag 1 (predicting flow rate for 1 day in advance). Mean Absolute error was 10.06733 and Median Absolute error was 4.34 when we used RandomForestRegressor Algorithms with 100 estimators in a Regression Model. While with XGBRegressor it was 10.37388 and 4.54288 respectively.

# **CONCLUSION**:

This solution could be used by the City of Calgary in forecasting the flood events along with the Bow River. This could be used to warn citizens of danger, and to put in place mitigation strategies to lessen the damages resulting from the flood. Other groups could also use the model for predicting daily flow rates along the Bow River. This could include recreational organizations involved in fishing along the Bow River or whitewater paddling in Harvie’s Passage.

Lessons learned while developing the model were:

* Finding resourceful information about natural resources from the Government websites were convenient and informative.
* After finding the information, reformatting data was quite time-consuming and needed few coding to merge different .csv files. Also, I learnt during that process that when data is not readymade, you can learn that which data will be helpful and informative for ongoing process, which data is useless, how can you handle unwanted and wanted data.
* Predicting nearly accurate flow rate on happened event was a difficult task. You must carefully handle data when wrangling & pre-processing. Moreover, our most important features might be different when there is different data was given or the value of data is changed in different events. If you predict the same data for time lag 1 & time lag 2 then the results are different for the same days.

Solution can be improved by:

* If we could have found more input features such as water use & demand, rain & snow, soil moisture reading and many more then it could have impact on predicting flow rate hence flood events.

# 