POURIA RAMAZI, Univesrity of Alberta
Predicting Infestations Using Machine Learning

Predicting future infestations is crucial to forest management. The many successes of machine learning across a variety of tasks, have motivated us to explore this approach for this important task. There are, however, many subtle challenges here: how to identify the best learning models, the best subset of the available covariates, and importantly, how to properly evaluate the models to avoid misleading performance measures (which are unfortunately common). We explore these ideas in the context of predicting the chance of a mountain pine beetle outbreak in the Cypress Hills area, over the years of 2006–2018, seeking the learned models with the best performance at predicting 1 year (resp., 3, 5, 7 years) infestation. We have found a generalized boosted regression tree (GBM) predicting the future 1 and 3-year infestations with 97% and 89% AUC, a neural network (NN) with one hidden layer predicting future 5-year infestations with 88% AUC, and a novel mixed model predicting future 7-year infestations with 80% AUC. We also show how incorrect model evaluations can lead to wrong performance measures: If the train and test datasets are obtained from a random split of the original dataset (rather than the appropriate year-based split), a generalized linear model, for example, would score 77% instead of the more accurate 62%. We then investigate how the prediction accuracy varies with respect to the provided history length of the covariates, and find that GVM, NN, and naive Bayes predict more accurately as the history length increases.