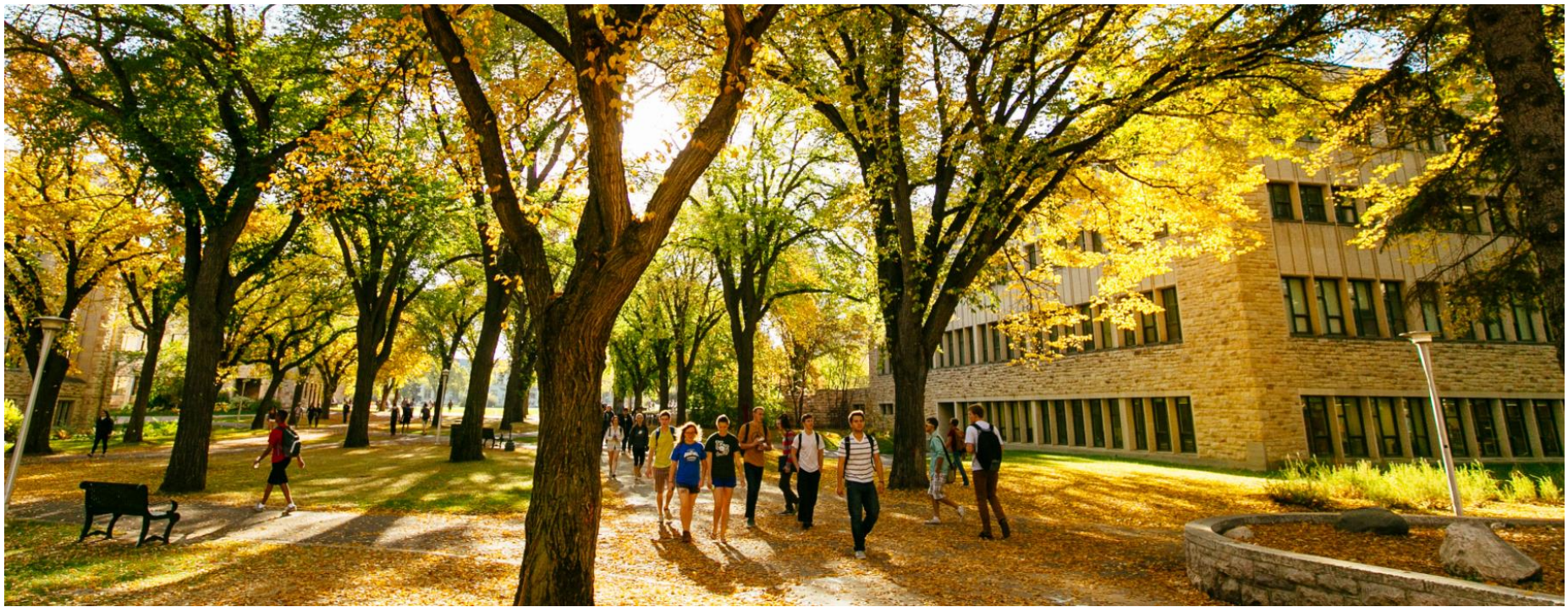




Twitter Data for Influenza Detection and Surveillance in Saskatchewan

Yuan Tian

PhD Candidate, Department of Computer Science



Background

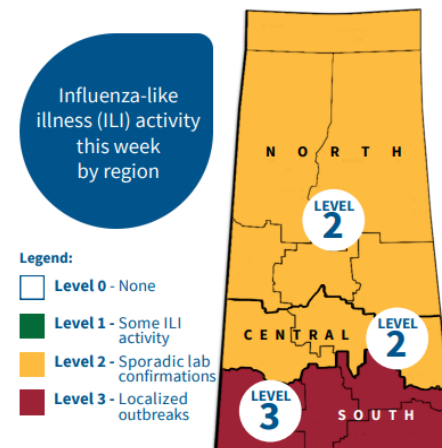
- Influenza infections caused **substantial mortality**
 - *250,000 – 500,000* deaths worldwide per year
- Influenza **surveillance** systems (such as FluWatch)
 - **Laboratory** reports of positive influenza cases
- In the era of “Big Data”, an enormous volume of **user-generated** data was produced and offers **new options** to **expand** current influenza surveillance.



Objective

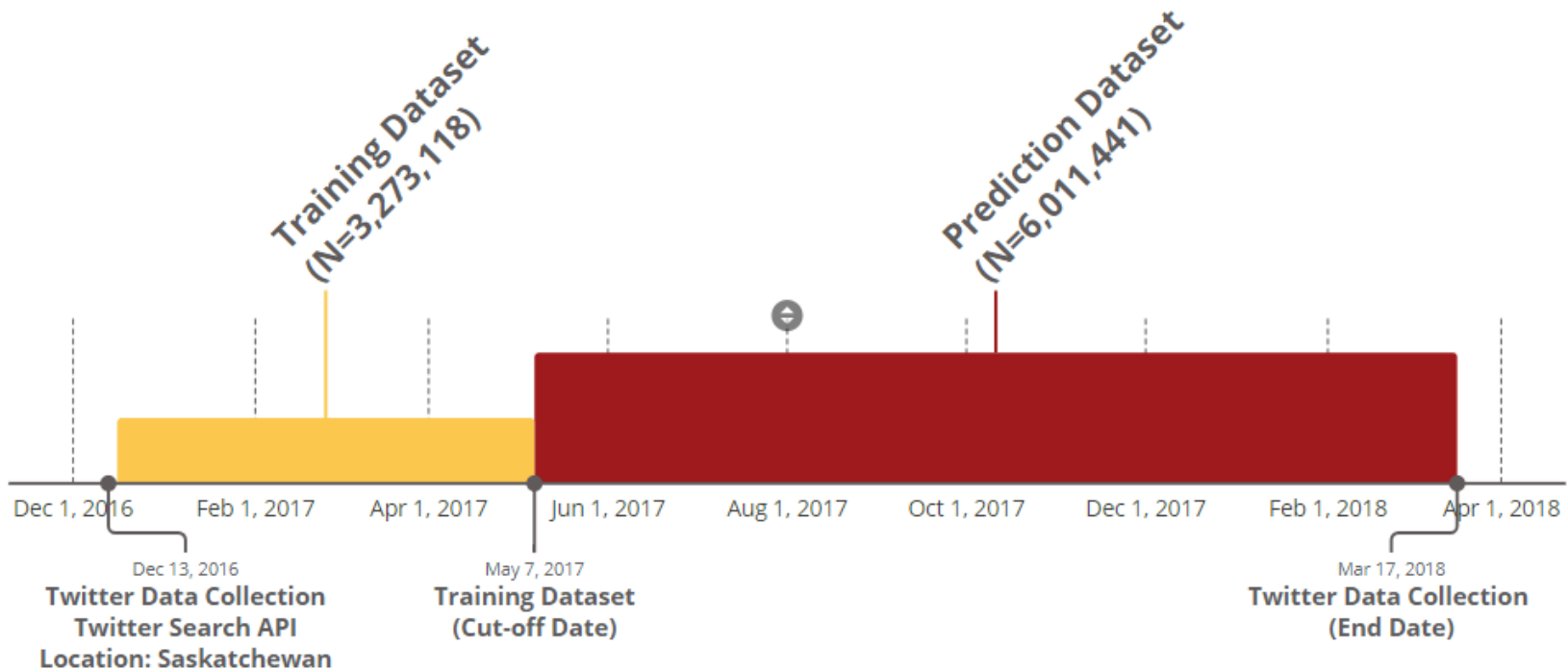
- explore whether the **trained classification models** equipped with **Twitter** data could complement **influenza surveillance** in the Canadian province of Saskatchewan
- Compare **performance** of different classification models.

Twitter Inc.
Social network company



Method

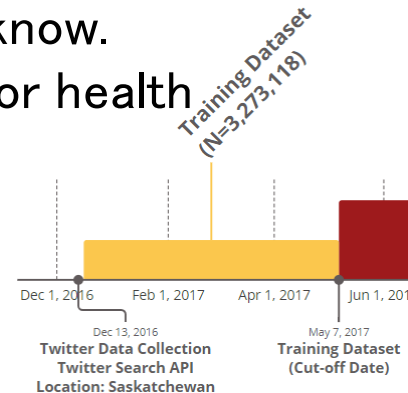
- **Twitter** Data Collection and Extraction
 - Over 12 million tweets in Saskatchewan
 - We excluded 2.7 million non-English or duplicated tweets.



Method

Training Dataset

- **Label** tweets as to whether they suggest **plausible influenza cases** (If yes, then 1; else then 0).
 - Step 1. Filter influenza-related tweets with **Keywords**
 - **Broader** inclusive criteria
 - Surveillance case definitions
 - Terms related to **Flu-like symptoms** / **influenza-like illness**
 - Step 2. **Manually** label tweets obtained from Step 1.
 - It could implicate **the user or someone** the user might know.
 - Tweets about flu-related advertisements, media news, or health promotion will be **labeled 0**.



Step 1. Filter influenza-related tweets with **Keywords**

Examples of Tweets **Implying** Influenza Cases:

- *“Home with a **cold/flu** surrounded by Kleenex; blankets and some light reading material”*
- *“Got a visit in last night with this sick little boy battling some **flu like symptoms**”*

Examples of Tweets **Not** Implying Influenza Cases:

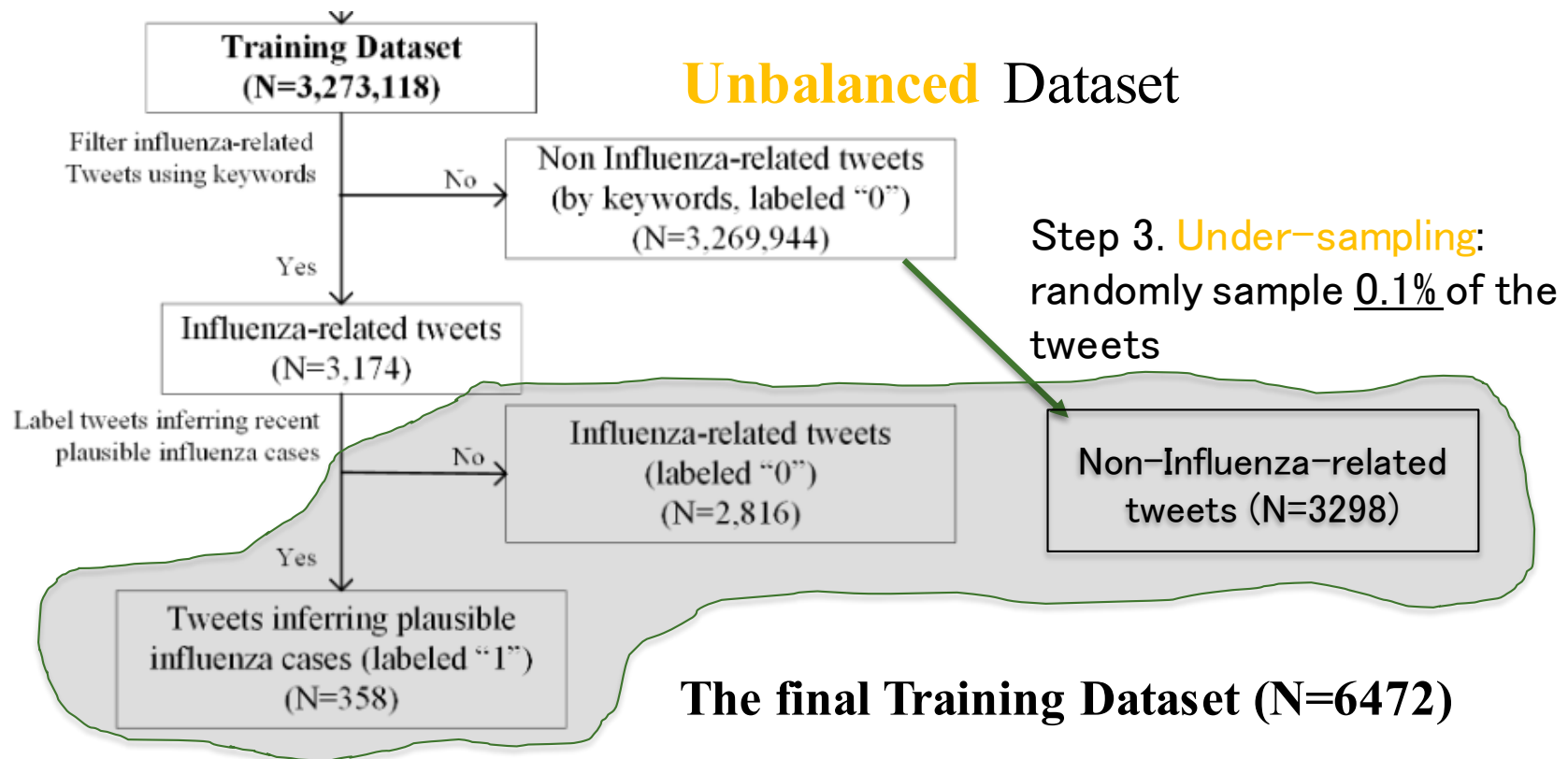
- *“Get **Your Flu Fighter Juice**.. Freshly squeezed carrots; apple; orange ginger.. Packed full of essential vitamins”*
- *“@rqhealth: **In the RQHR you can get a flu shot** by making an appointment at a Public Health Office.”*



Method

Training Dataset

- Step 1. Filter with **Keywords**
- Step 2. **Manually** label tweets



Method

Training Dataset

- Tweets inferring plausible influenza cases (N=358)

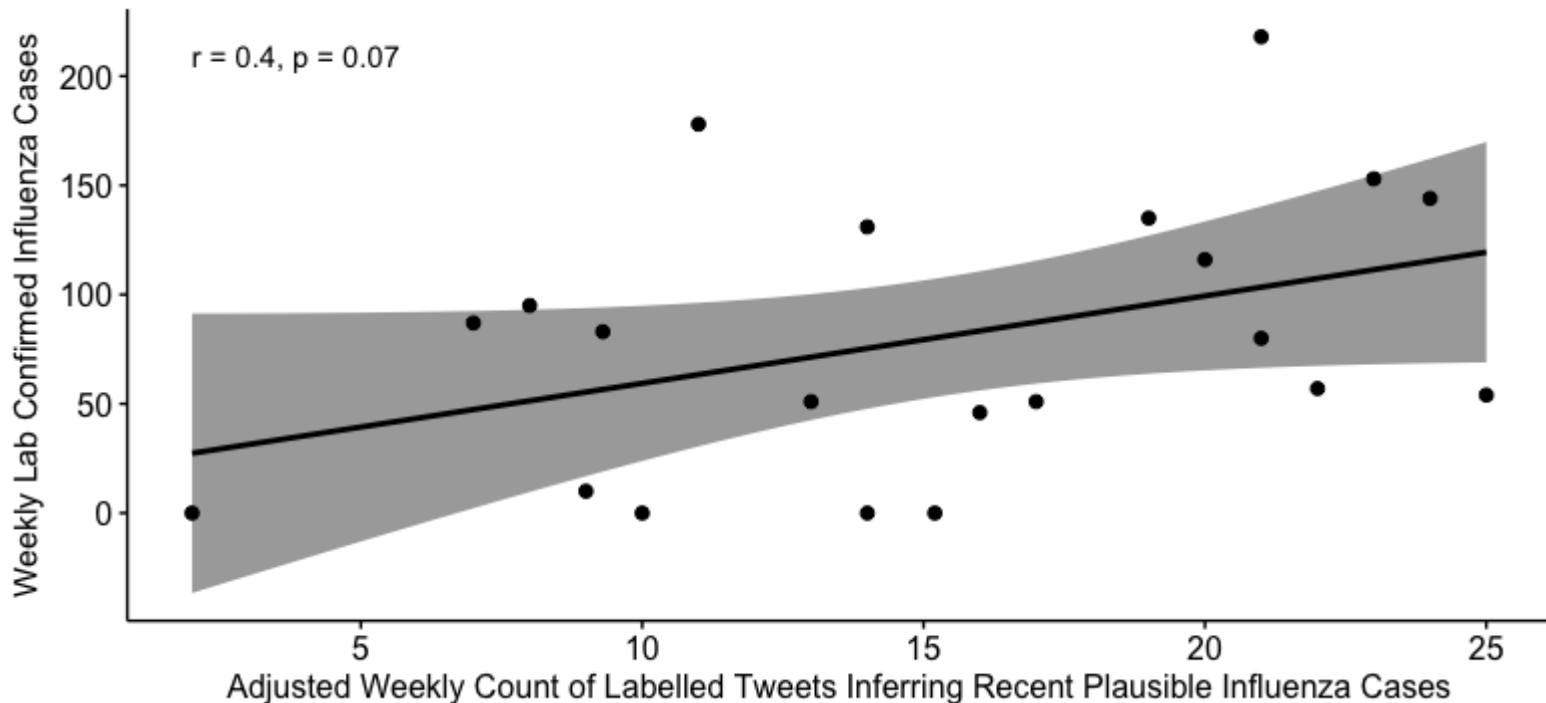
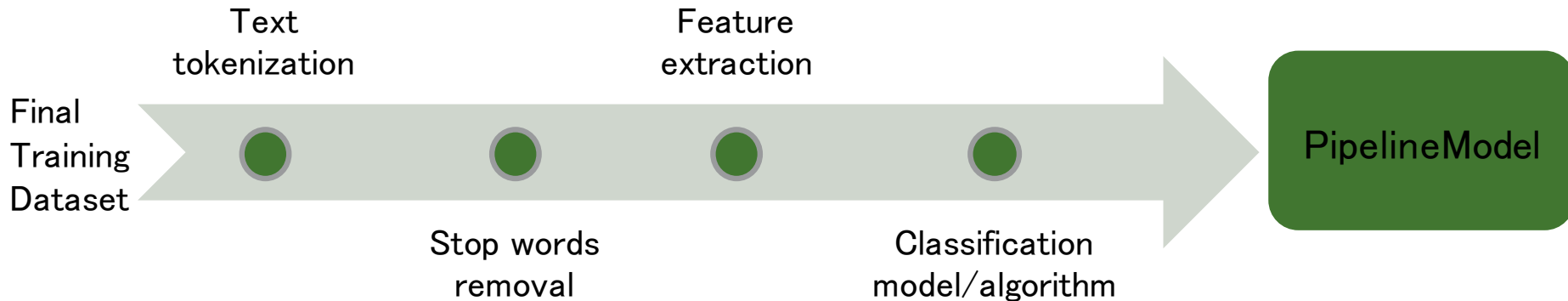


Figure. Scatter-plot of Saskatchewan Weekly Lab Confirmed Influenza Cases and Adjusted Weekly Counts of Tweets Suggesting Recent Plausible Influenza Cases between 2016-12-13 and 2017-05-07

Method

Classification Models

- Logistic regression model
- Decision tree classifier
- Random forest classifier



** 3-fold cross-validation was performed for each model for evaluation purpose*



Results – Model Comparison

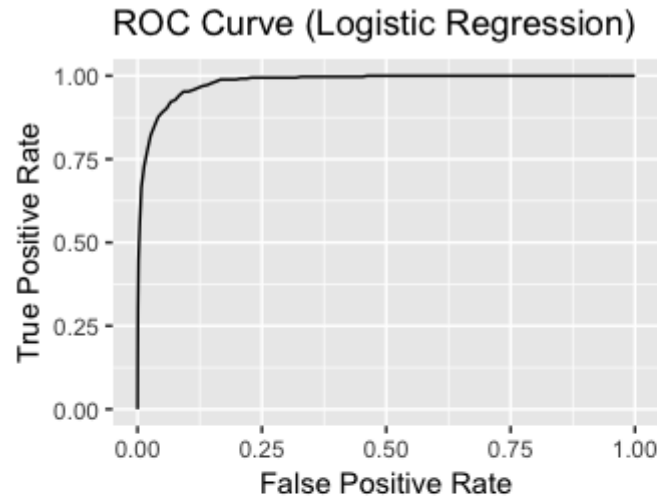
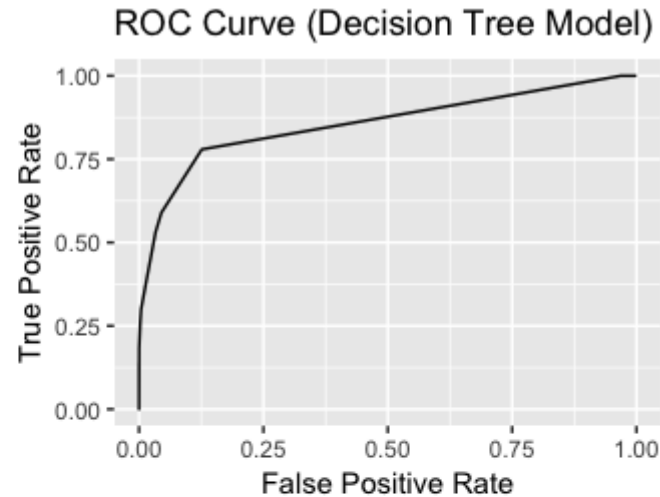
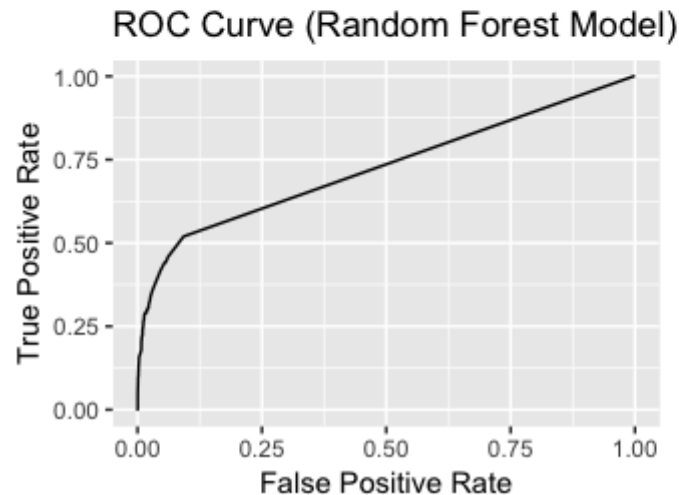


Figure. Receiver Operating Characteristics (ROC) Curves for three different classification models. Area under ROC curve for Logistic Regression (0.981), Decision Tree (0.856) and Random Forest Model (0.726)



Results – Model Comparison

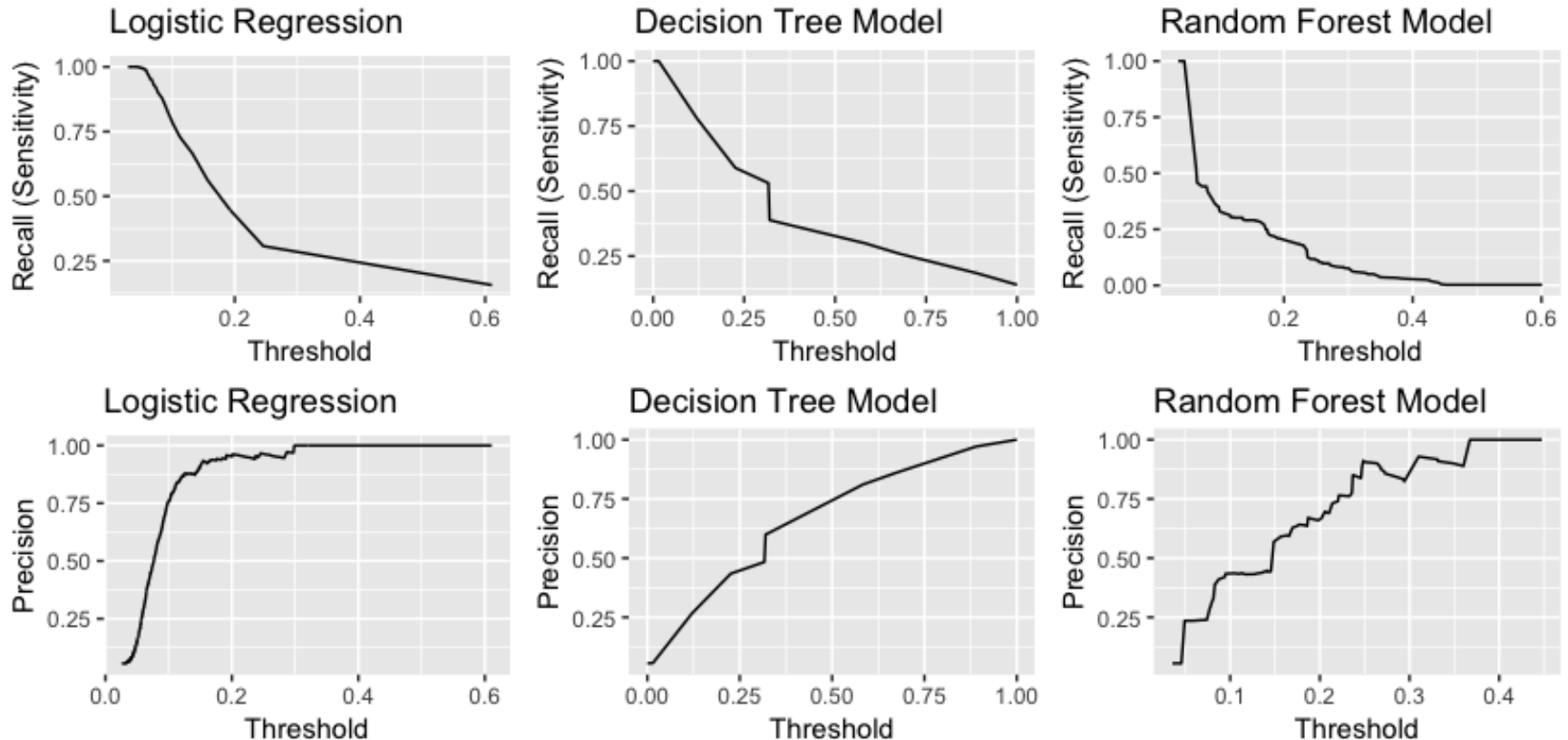


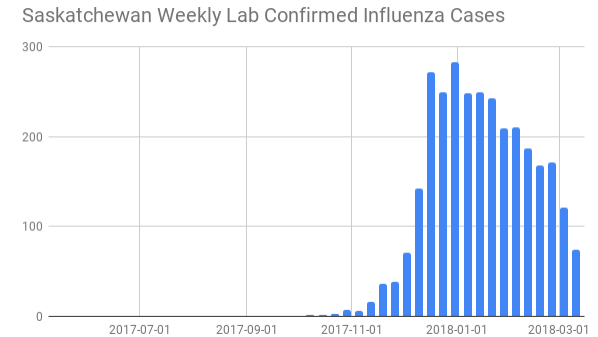
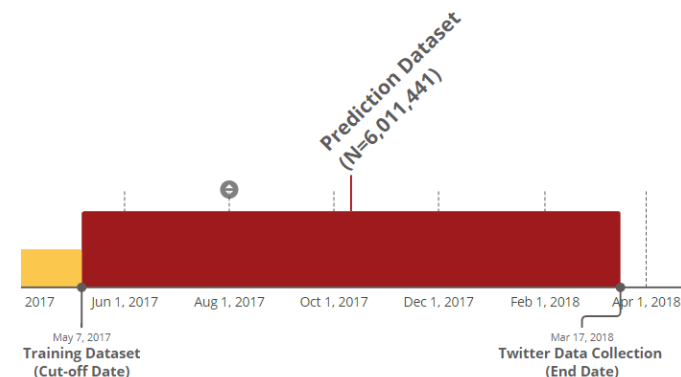
Figure. Recall and Precision



Results – Validation

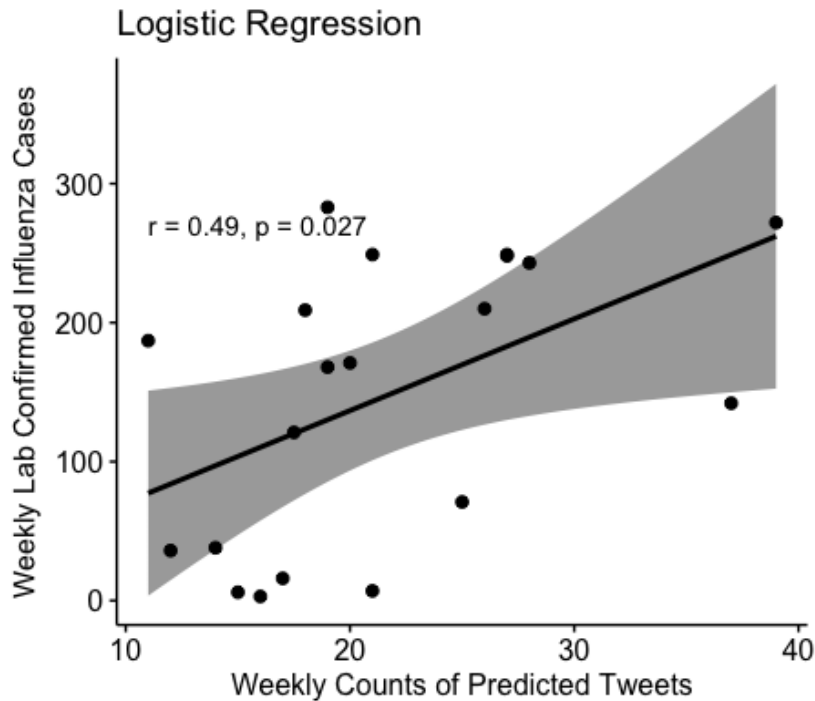
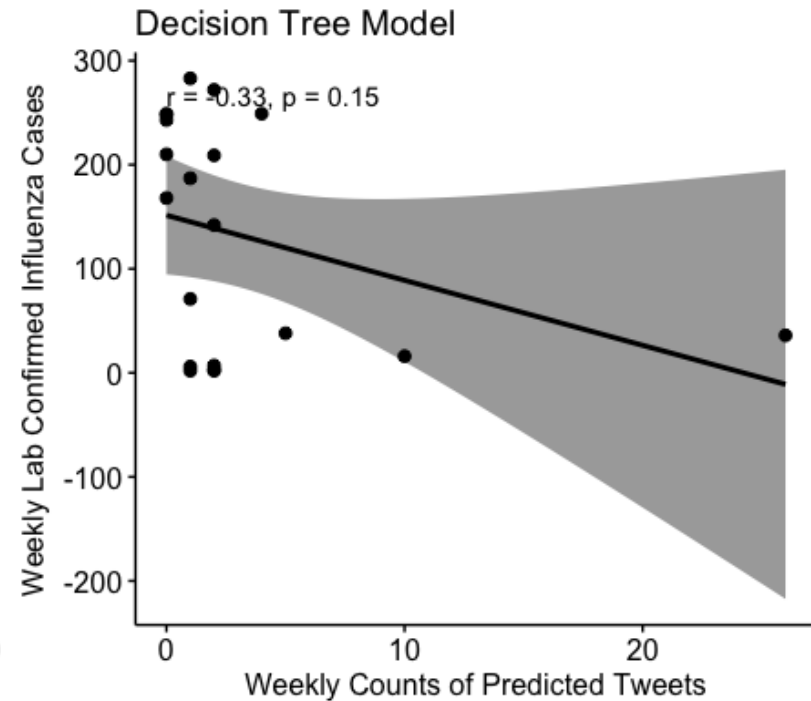
Out-of-sample Prediction

- The prediction dataset was filtered and under-sampled (N=12449).
- Predict using the trained classification models (minimum recall = 0.75)
- Aggregate the predictions into weekly counts and assess their correlation with Saskatchewan weekly lab confirmed influenza cases.



Results

Out-of-sample Prediction (Validation)



There is a **strong** correlation between lab confirmed influenza cases and predictions made by the fitted logistic regression model ($p=0.027$)



Conclusion

- Trained classification models equipped with Twitter data has the potential to complement influenza surveillance in Saskatchewan
- The performance of different classification models varies.



Limitation

- Manual labelling
- Laboratory reports of influenza cases



Thank you

