# Are you open?

Naren Pandian(npandian), Vaibhav Aggarwal(vaibhavg)

## Abstract

*In this paper we propose a new machine learning approach to predicting whether a business is open or permanently shutdown. This problem is of keen importance because with the dynamic world the information available online becomes stale very fast. The impact of serving stale listing is lost trust. On the other hand it is cost prohibitive to manually verify every business listing for its freshness. Hence we propose using a learning algorithm to classify businesses as shutdown with high probability. We use the Yelp dataset for training and testing. We start with the standard techniques of Logistic Regression and SVM, and then propose our own algorithms EMSVM and EMLOG. Instead of using the traditional RMSE metric we use the more expression Precision-Recall metric to measure performance. This metric allowed us to prioritize one at the expense of the other to fit over use case.*

## Dataset

## Feature set

**Assessment Feature**
- Number of reviews for a business.
- Average rating for the business.
- Number of user checkins.

**Temporal Features**
- Number of days since last review.
- Number of days since last tip.

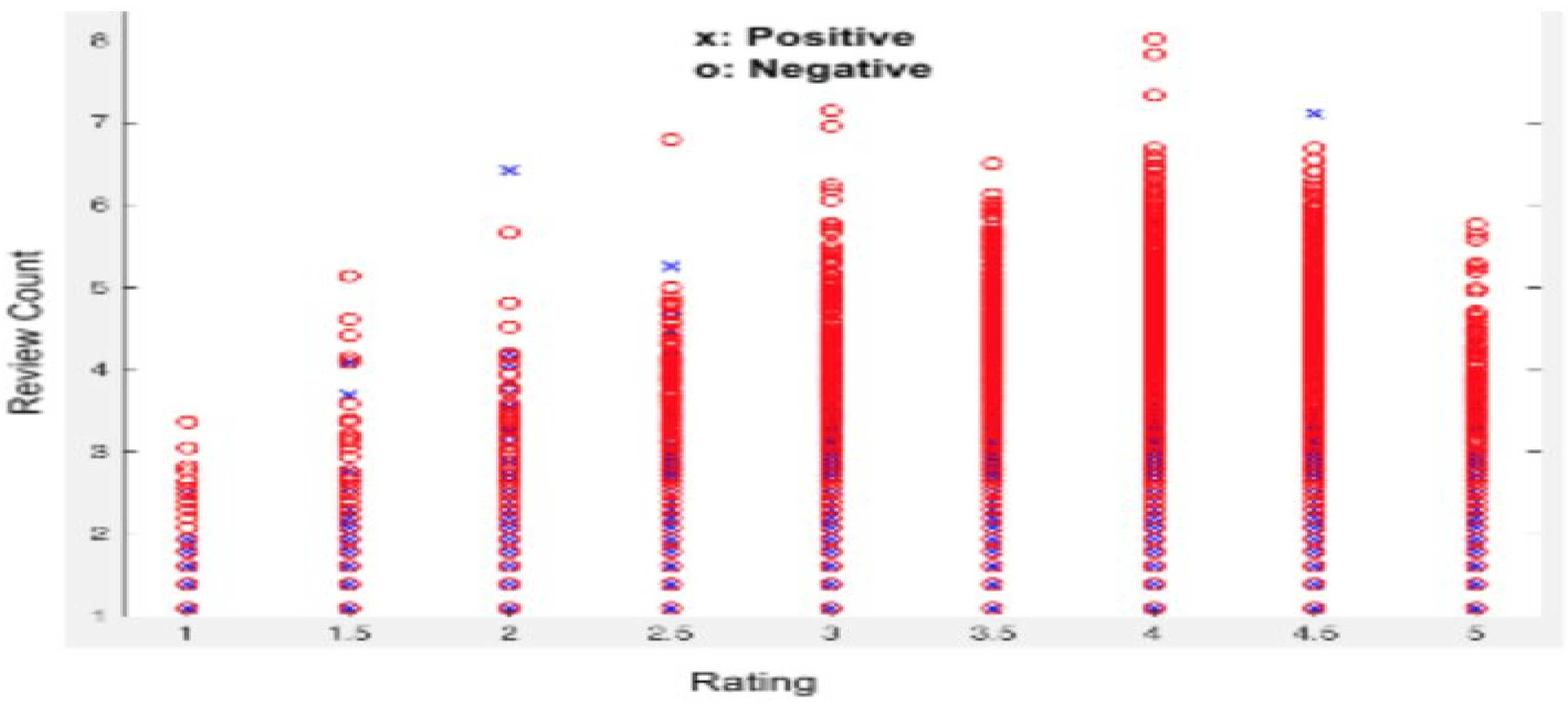
**Spatial Features**
- Latitude.
- Longitude.
- State.

**Vocabulary Features**
- Tokenized review text.

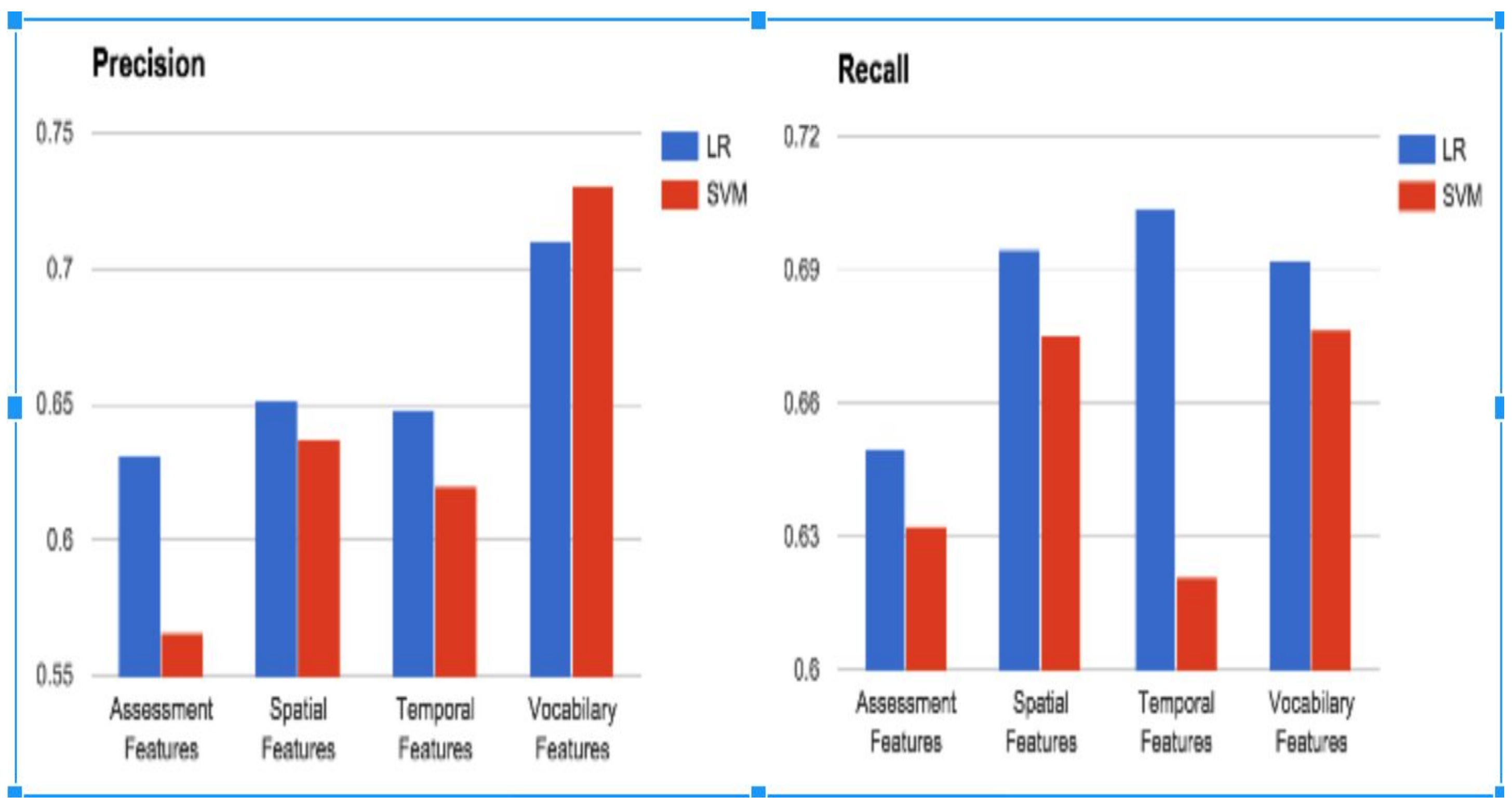
## Baseline with SVM

| Scenarios | Precision | Recall |
|---|---|---|
| Raw data | 0.057143 | 0.063291 |
| Unskewed data | 0.49819 | 0.7759 |
| Raw data, k-fold cv, k=3 | 0.12445 | 0.092589 |
| Unskewed data, k-fold cv, k=3 | 0.48662 | 0.5698 |

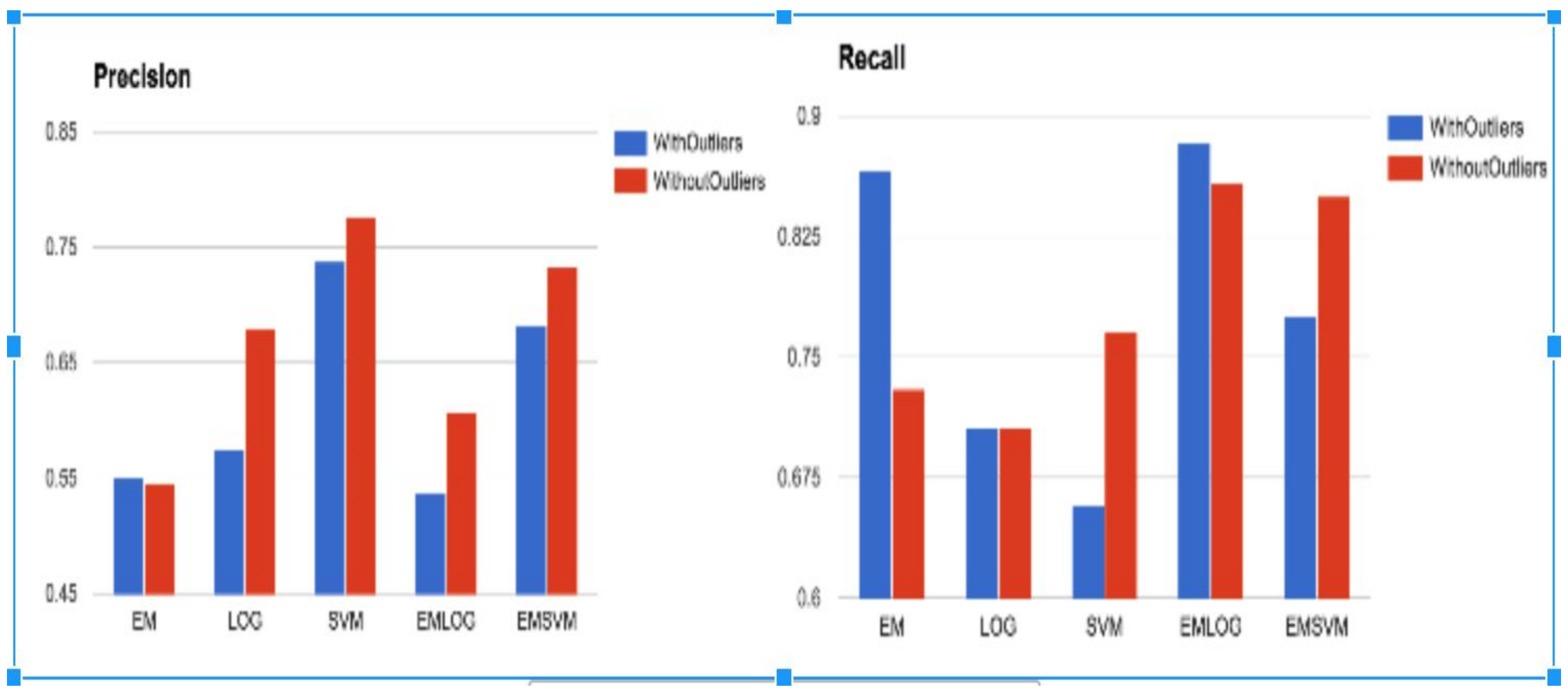
## Data Visualization



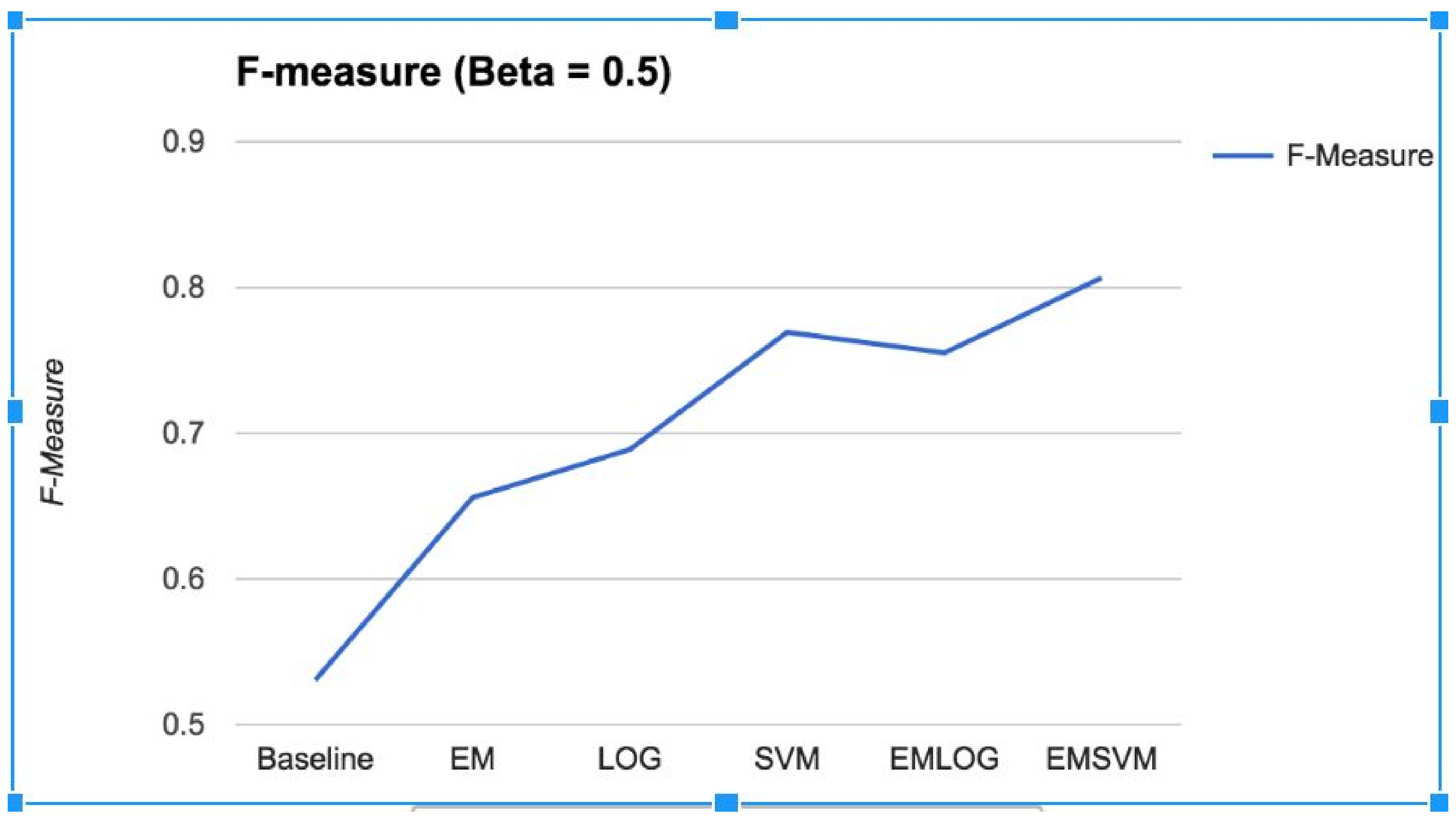
## Feature Improvement



## Data Distribution Study



## Algorithimic Improvement



## Summary

- **Inherent bias problem with basic features**
- **Improvement in precision and recall after adding new features from the dataset.**
- **Studied the impact of skewness and the outliers in the data set.**
- **We are proposing our own EMLOG and EMSVM which ensembles EM clustering with Logistic Regression and SVM. They provide much higher recall at slight expense of precision compared to traditional algorithms**