# StumbleUpon

## Classifying Evergreen vs Ephemeral Content

Riwa Sabri

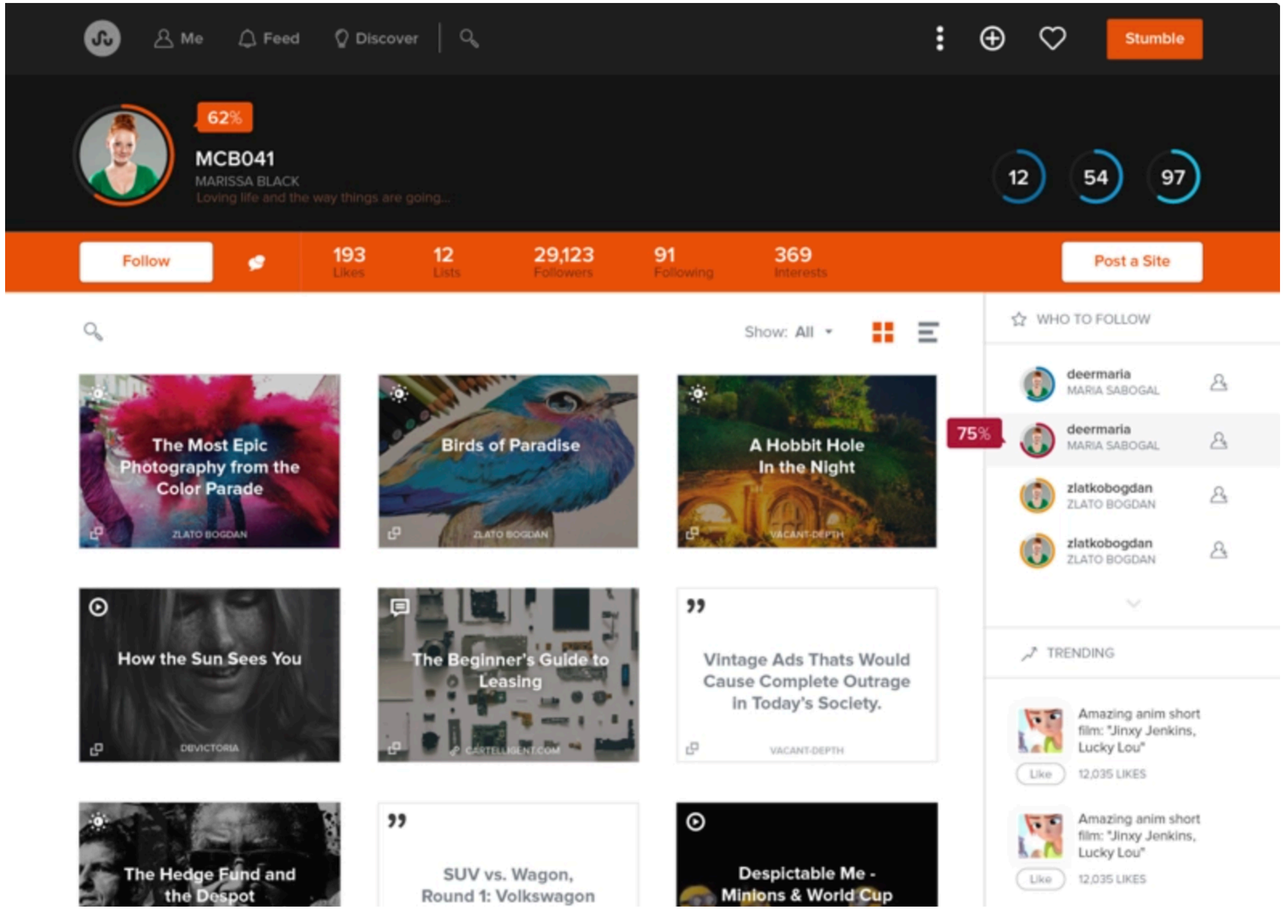
October 2022

SECTION

# Background

# About

## What is StumbleUpon, now Mix?




StumbleUpon is a user-curated web content discovery engine, which recommends quality pages and media to its users.

While some recommended content like news or seasonal content is not relevant over a long period of time - Ephemeral - some content can be relevant for years - Evergreen.

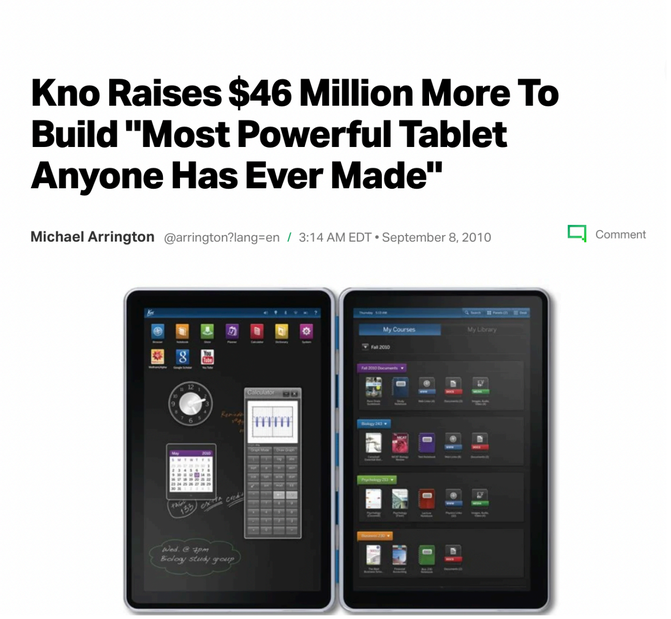
# Scope

## Classifying Evergreen vs Ephemeral Content to improve User Experience

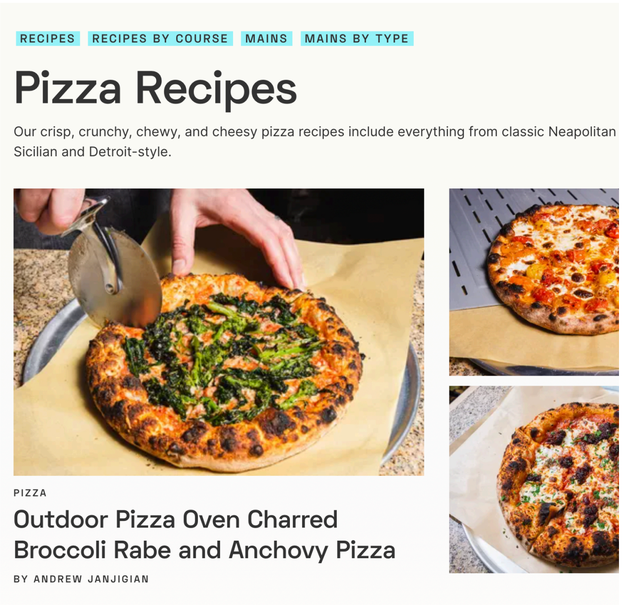
- StumbleUpon gets rating from their community that allows them to assess whether a page is relevant.

- However, they would like to be able to **make this distinction ahead of time** by building a classifier that can label media as evergreen or not.

*Example of recommended media classified by users*




Rated Ephemeral




Rated Evergreen

# Methodology

## Process

**EDA**

Data Cleaning and Visualization to better understand the problem and metrics of interest.

01

02

**Baseline Modeling**

After feature selection using Lasso on Logistic Regression initial model, several baseline models of varying complexity were tested on these features and tuned using RandomSearchCV and GridSearchCV.

**Feature Engineering**

Leveraging industry insights to extract relevant categorical features from the dataset and engineer new features. Re-evaluation and tuning of models of varying complexity with new features.

03

04

**More Feature Engineering with TF-IDF**

Testing out another approach relying on articles content.

# Methodology

## Metrics of interest

### Precision

$$\frac{\text{True Evergreen}}{\text{Classified Evergreen}}$$

Precision will allow me to reduce the number of false positives. This will decrease the likelihood of misclassifying ephemeral as evergreen.

### Accuracy

$$\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Accuracy is informative here because the classes evergreen/ephemeral are balanced in the dataset.

SECTION

# Model Iteration

Methodology Note

# 5-fold cross validation scores are used for model comparison

PART I

# *Baseline Models*

The following models consist of models of varying complexity trained on a select number of numerical features

# Selecting features for the baseline model

## Features Description

The following numerical features were selected for the baseline model:

| Feature Name | Description |
|---|---|
| linkwordscore | Percentage of words on the page that are in hyperlink's text |
| frameTagRatio | Ratio of iframe markups over total number of markups |
| non_markup_alphanum_characters | Page's text's number of alphanumeric characters |
| commonlinkratio_2 | # of links sharing at least 1 word with 2 other links / # of links |
| commonlinkratio_3 | # of links sharing at least 1 word with 3 other links / # of links |
| commonlinkratio_4 | # of links sharing at least 1 word with 4 other links / # of links |
| numwords_in_url | Number of words in url |
| avglinksize | Average number of words in each link |
| html_ratio | Ratio of tags vs text in the page |
| compression_ratio | Compression achieved on this page via gzip (measure of redundancy) |
| numerofLinks | Number of markups |
| is_news | True (1) if StumbleUpon's news classifier determines that this webpage is news |
| image_ratio | Ratio of  tags vs text in the page |

# The baseline model yields accuracy & precision around 60%

## Numerical Features

**Our baseline logistic regression with L2 Penalty model containing only our raw base features yields accuracy and precision close to 60%**

| Accuracy | Precision |
|---|---|
| 62.07% | 61.24% |

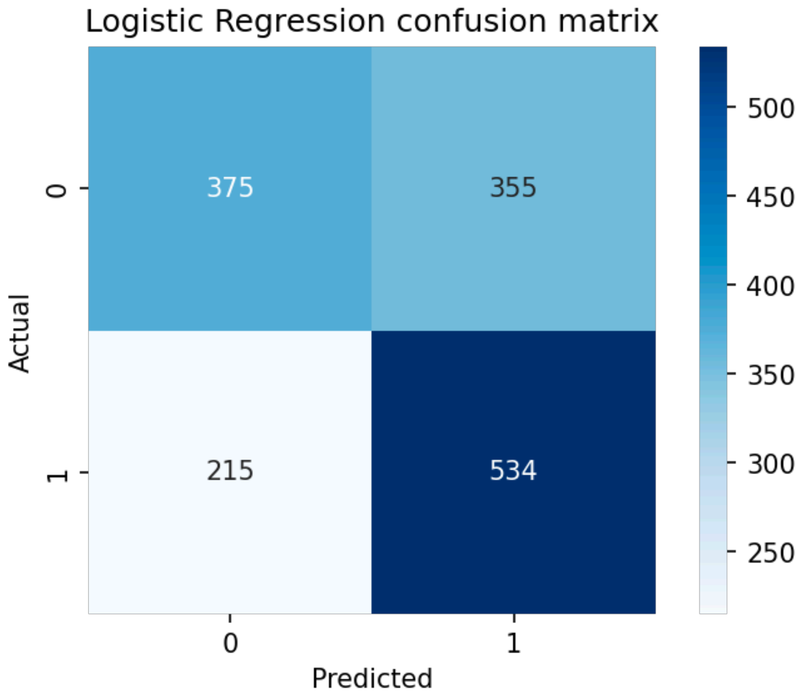
# And is better at classifying positives than negatives

## Numerical Features

**Our baseline logistic regression model containing only our raw base features yields accuracy and precision close to 60%**

| Accuracy | Precision |
|---|---|
| 62.07% | 61.24% |

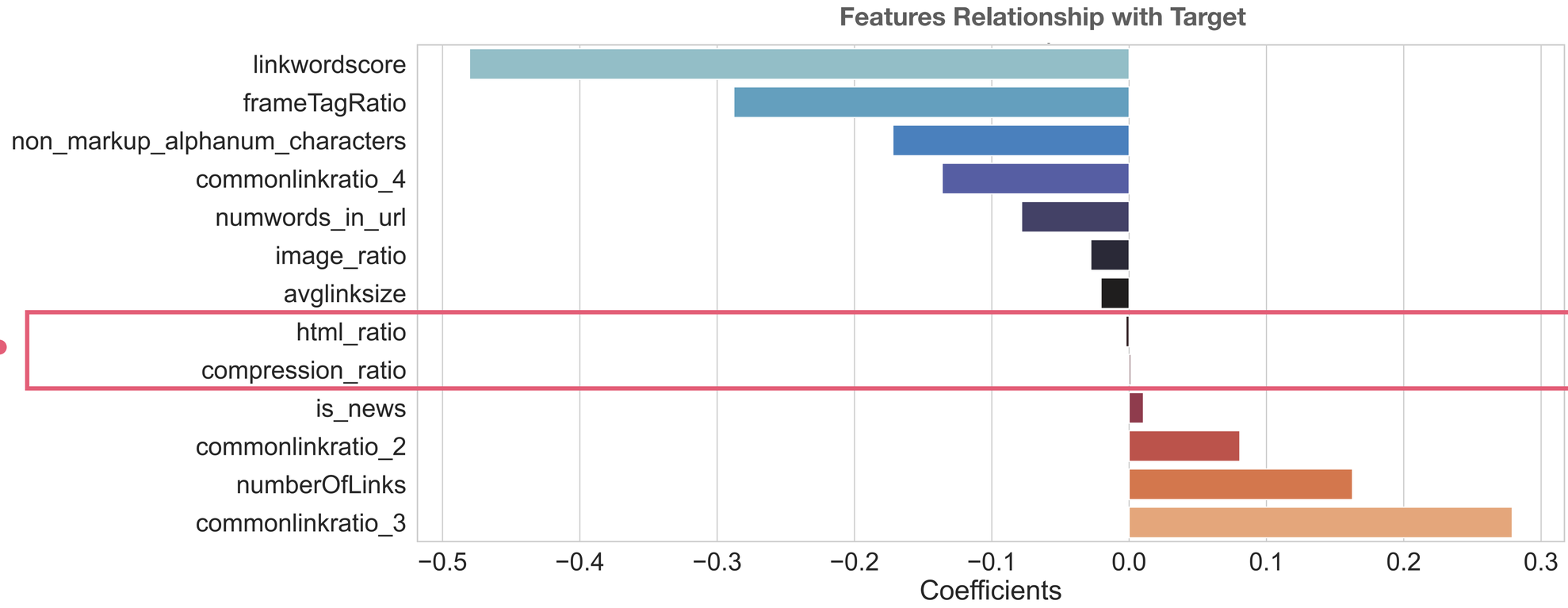
**The baseline model is better at classifying positives than it is at classifying negatives.**



# Eliminating features with coefficients close to 0 with L2

## Re-evaluating Feature Selection

html ratio & compression ratio have really small coefficients and will be removed



# Comparing baseline Models of varying Complexity

## Results: Random Forest has the best accuracy and precision

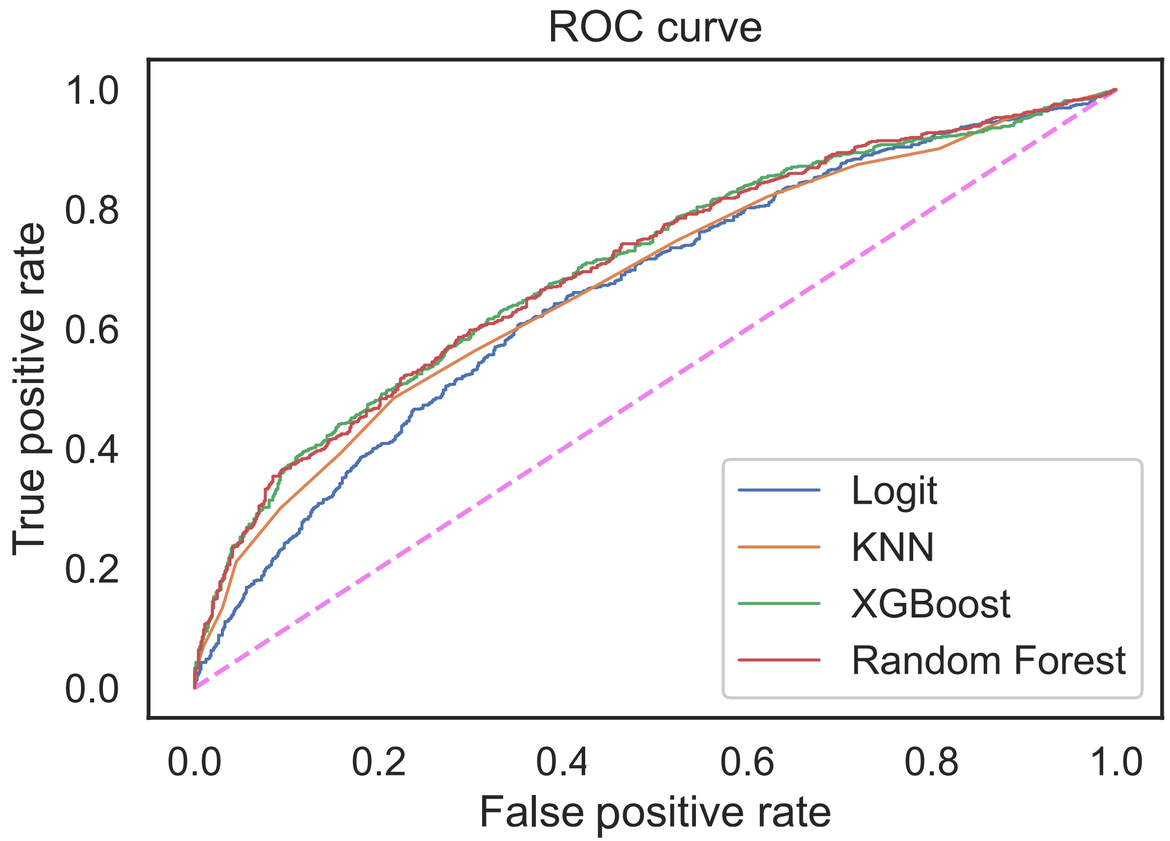
*Note: I used RandomSearch CV/GridSearchCV for hyper parameter tuning for all models*

| | Accuracy | Precision |
|---|---|---|
| **Logistic Regression** | 62.22% | 61.40% |
| **KNN** | 63.73% | 64.38% |
| **Decision Trees** | 59.60% | 60.55% |
| Best Baseline → **Random Forest** | 65.67% | 65.55% |
| **XGBoost** | 64.96% | 64.89% |

# Random Forest has the highest AUC

## ROC AUC Comparaison

*Note: I used RandomSearch CV/GridSearchCV for hyper parameter tuning for all models*



PART II

# *Feature Engineering*

The following section consists of engineered features and their effect on model performance.

October 2022 | Riwa Sabri

# Making the data more informative by extracting relevant categorical variables that were previously unused

| Feature | Description | Examples |
|---|---|---|
| Publisher | Publisher of the article | Men’s health, Bloomberg |
| Alchemy_Category | Overall topic of the article | Recreation, Sports |
| News_Frontpage | True if StumbleUpon's news classifier determines that this webpage is front-page news | 1,0,? |

# Engineering New Features by using Domain Knowledge

## Incorporating domain knowledge to the features (1/2)

### Insight and resulting feature for Evergreen Article Identification

| Insight | Feature |
|---|---|
| The industries that have the highest evergreen to ephemeral article rate are digital marketing, technology and health. | True if article's url contains one of these words. |
| Articles with the following content are more likely to be relevant over time: how to's, recipes. | True if article's url contains one of these words. |

# Engineering New Features by using Domain Knowledge

## Incorporating domain knowledge to the features (2/2)

### Insight and resulting feature for Ephemeral Article Identification

| Insight | Feature |
|---|---|
| The industries that have the highest ephemeral to evergreen article rate are: SEO, business and fashion | True if article's url contains one of these words. |
| Articles with the following content are less likely to be relevant over time: reports, statistics | True if article's url contains one of these words. |

# Results after Feature Engineering

## New features lead to improvements in precision and accuracy

| | Accuracy | (% change from RF Baseline) | Precision | (% change from RF Baseline) |
|---|---|---|---|---|
| **Logistic Regression** | 72.70% | **+11%** | 74.70% | **+14%** |
| **KNN** | 70.16% | **+7%** | 71.60% | **+11%** |
| **Random Forest** | 72.65% | **+11%** | 74.11% | **+13%** |
| **XGBoost** | 72.48% | **+11%** | 73.46% | **+12%** |
| **Ensembling (RF, XGB, KNN)** | 72.07% | **+10%** | 73.26% | **+12%** |

Best Model

# Vectorizing Article Content with TF-IDF

## More feature engineering

**TF-IDF (term frequency-inverse document frequency)** is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

***TF-IDF on Article Content***

'"A sign stands outside the International Business Machines Corp IBM Almaden Research Center campus in San Jose Califo
rnia Photographer Tony Avelar Bloomberg Buildings stand at the International Business Machines Corp IBM Almaden Rese
arch Center campus in the Santa Teresa Hills of San Jose California Photographer Tony Avelar Bloomberg By 2015 your m
obile phone will project a 3 D image of anyone who calls and your laptop will be powered by kinetic energy At least t
hat s what International Business Machines Corp sees in its crystal ball The predictions are part of an annual tradit
ion for the Armonk New York based company which surveys its 3 000 researchers to find five ideas expected to take roo
t in the next five years IBM the world s largest provider of computer services looks to Silicon Valley for input glea
ning many ideas from its Almaden research center in San Jose California Holographic conversations projected from mobi
le phones lead this year s list The predictions also include air breathing batteries computer programs that can tell
when and where traffic jams will take place environmental information generated by sensors in cars and phones and cit
ies powered by the heat thrown off by computer servers These are all stretch goals and that s good said Paul Saffo ma
naging director of foresight at the investment advisory firm Discern in San Francisco In an era when pessimism is the
new black a little dose of technological optimism is not a bad thing For IBM it s not just idle speculation The compa
ny is one of the few big corporations investing in long range research projects and it counts on innovation to fuel g
rowth Saffo said Not all of its predictions pan out though IBM was overly optimistic about the spread of speech techn
ology for instance When the ideas do lead to products they can have broad implications for society as well as IBM s b
ottom line he said Research Spending They have continued to do research when all the other grand research organizatio
ns are gone said Saffo who is also a consulting associate professor at Stanford University IBM invested 5 8 billion i
n research and development last year 6 1 percent of revenue While that s down from about 10 percent in the early 1990
s the company spends a bigger share on research than its computing rivals Hewlett Packard Co the top maker of persona
l computers spent 2 4 percent last year At Almaden scientists work on projects that don t always fit in with IBM s co
mputer business The lab s research includes efforts to develop an electric car battery that runs 500 miles on one cha
rge a filtration system for desalination and a program that shows changes in geographic data IBM rose 9 cents to 146
04 at 11 02 a m in New York Stock Exchange composite trading The stock had gained 11 percent this year before today C
itizen Science The list is meant to give a window into the company s innovation engine said Josephine Cheng a vice pr
esident at IBM s Almaden lab All this demonstrates a real culture of innovation at IBM and willingness to devote itse
lf to solving some of the world s biggest problems she said Many of the predictions are based on projects that IBM ha

# Results show improvement in accuracy and precision

## Logistic Regression with TF-IDF

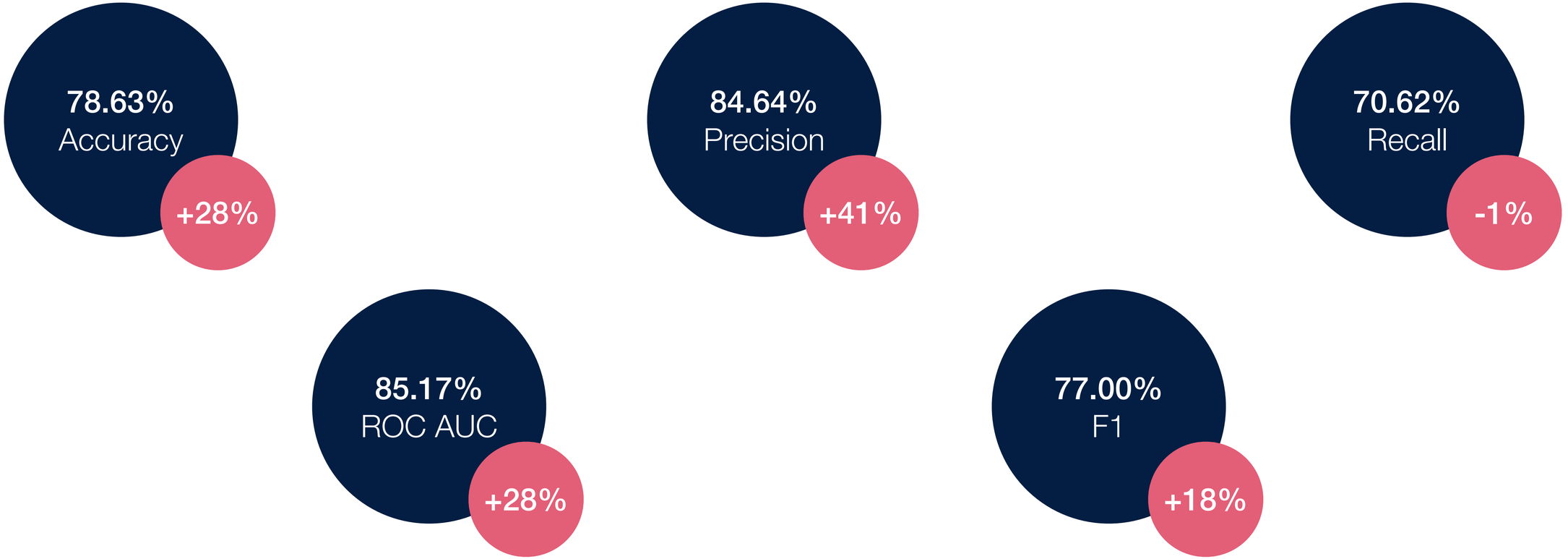
| | Accuracy | Precision |
|---|---|---|
| **Logistic Regression without TF-IDF** | 72.70% | 74.70% |
| **Logistic Regression with TF-IDF** | 78.10% | 84.57% |
| **(% change)** | **7% increase** | **13% increase** |

SECTION

# Conclusions & Future Work

# Baseline vs TF-IDF Model: Metrics on Test

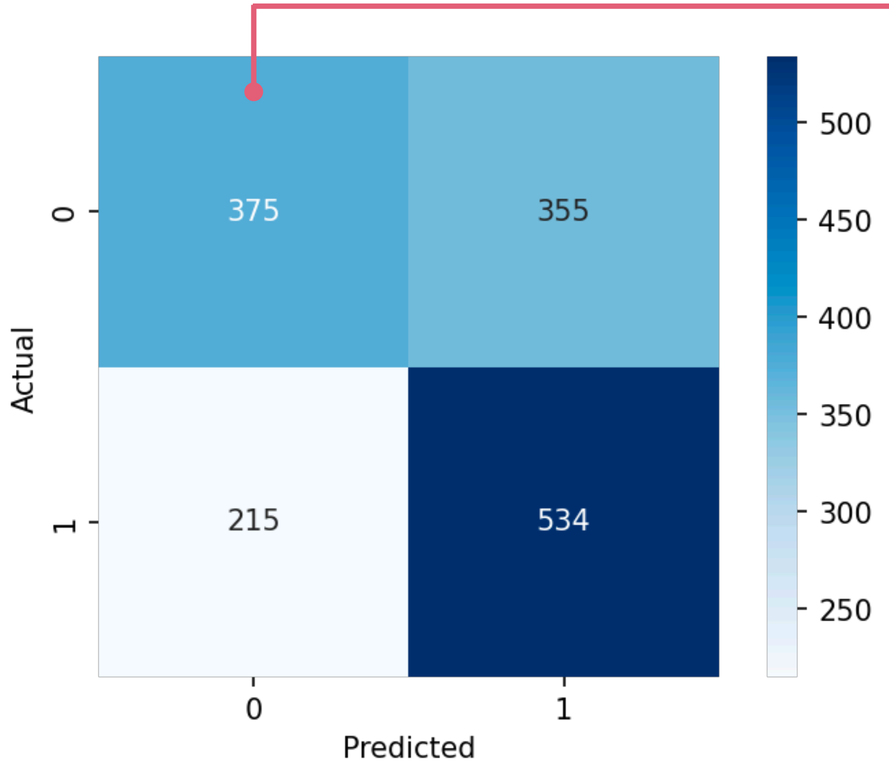
## Conclusions - Model Evolution



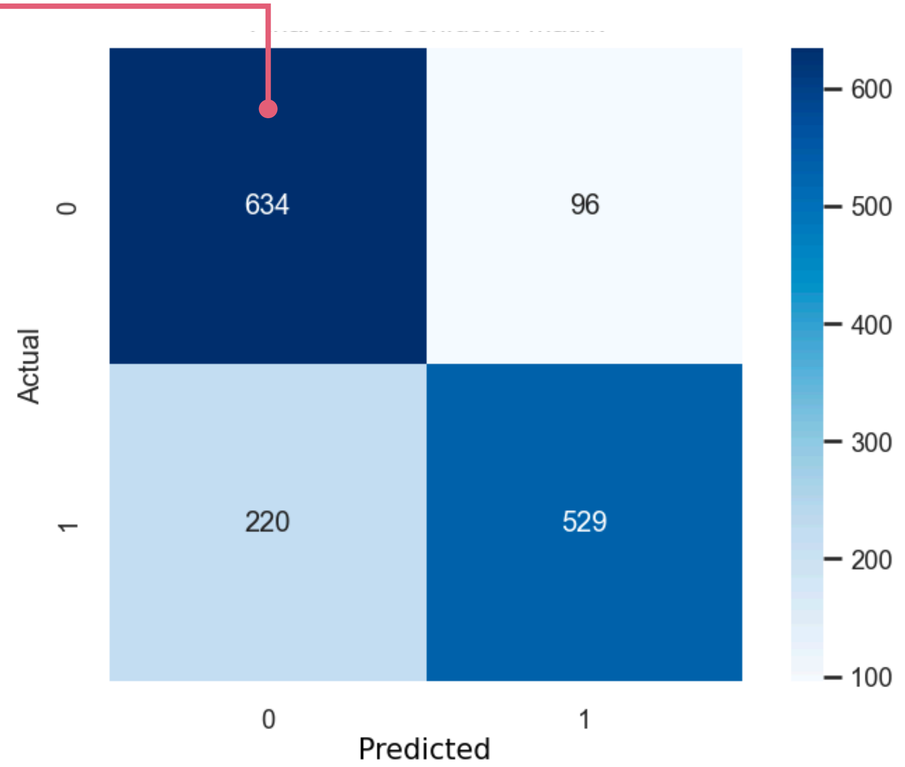
# Baseline vs TF-IDF Model: Confusion Matrix

## Conclusions - Model Evolution

Improvement in true negatives




*Baseline Model Confusion Matrix*

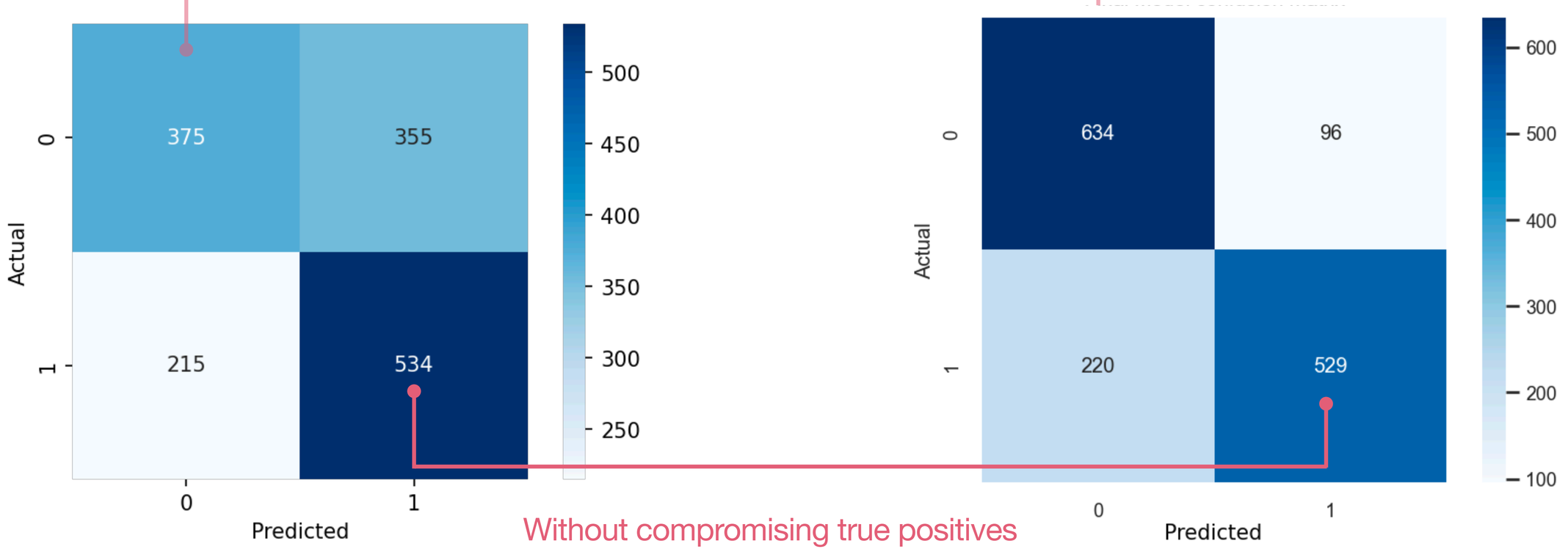



*Final Model (TF-IDF LogReg) Confusion Matrix*

# Baseline vs TF-IDF Model: Confusion Matrix

## Conclusions - Model Evolution

Improvement in true negatives




Without compromising true positives

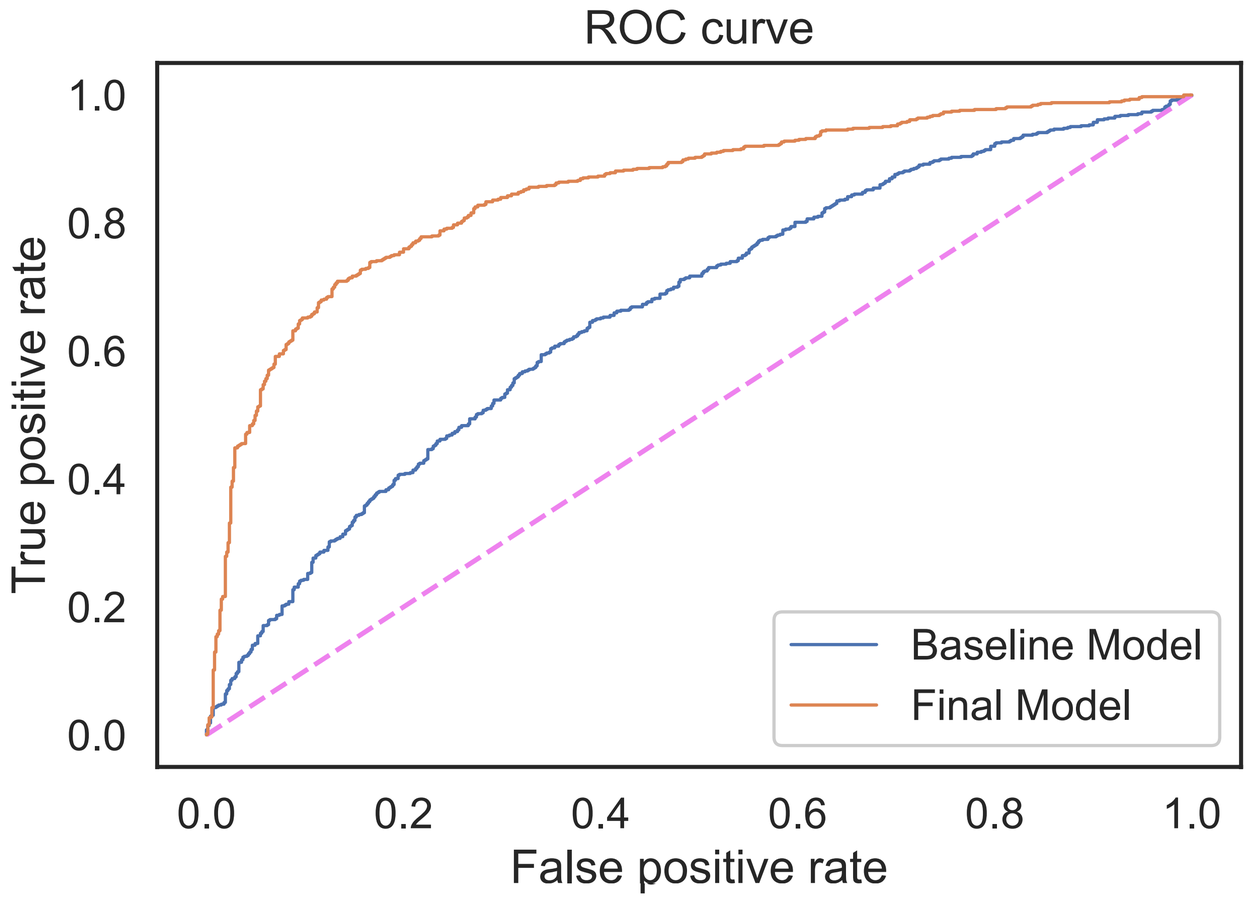



*Baseline Model Confusion Matrix*

*Final Model (TF-IDF LogReg) Confusion Matrix*

# Baseline vs TF-IDF Model : ROC_AUC

## Conclusions - Model Evolution



# Next Steps: Towards a text based solution

## Future Work

- Try TF-IDF on Article Title to see if it leads to improvement.

- Test more models with the last set of features including TF-IDF.

- Improve on feature selection using more rigorous techniques (chi square test, mutual information)

# Thank you