Capstone Project

CREDIT CARD DEFAULT PREDICTION

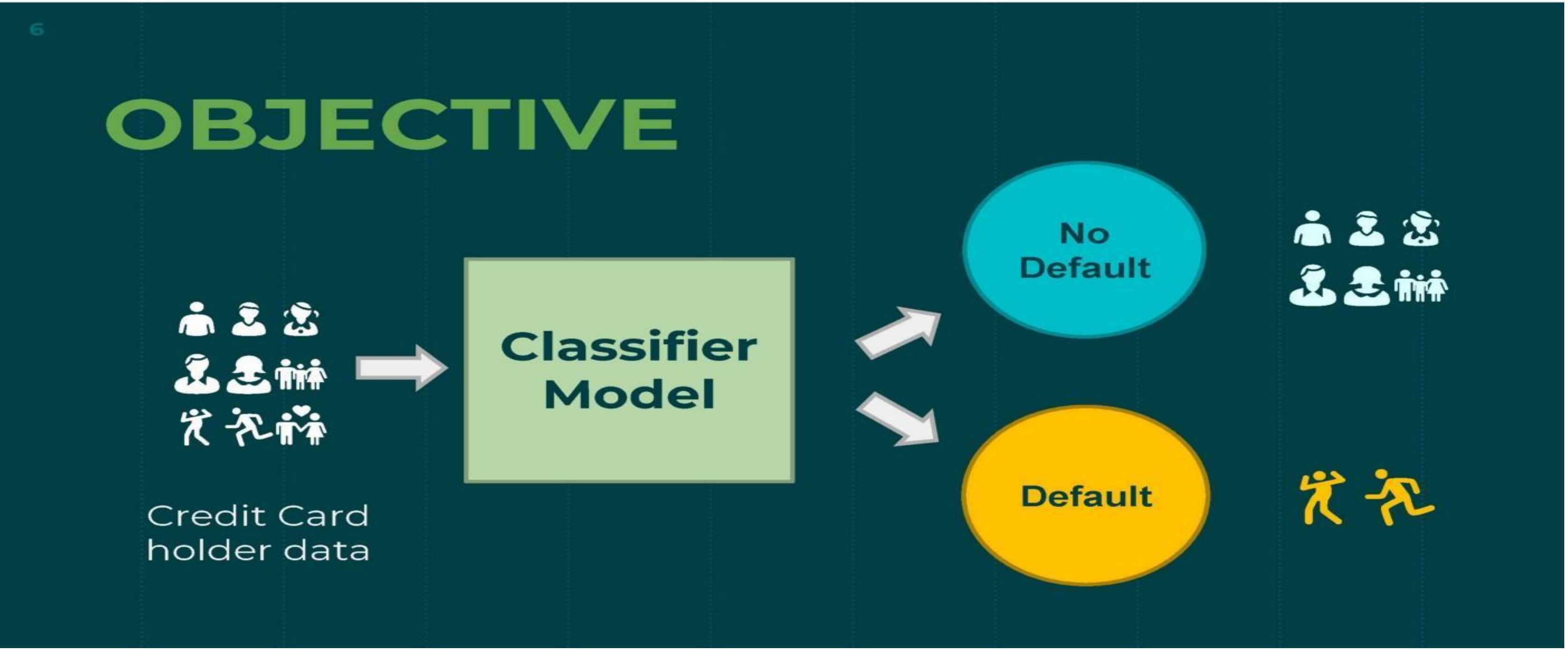
By – RAHUL SINGH WALDIA

CREDIT CARD

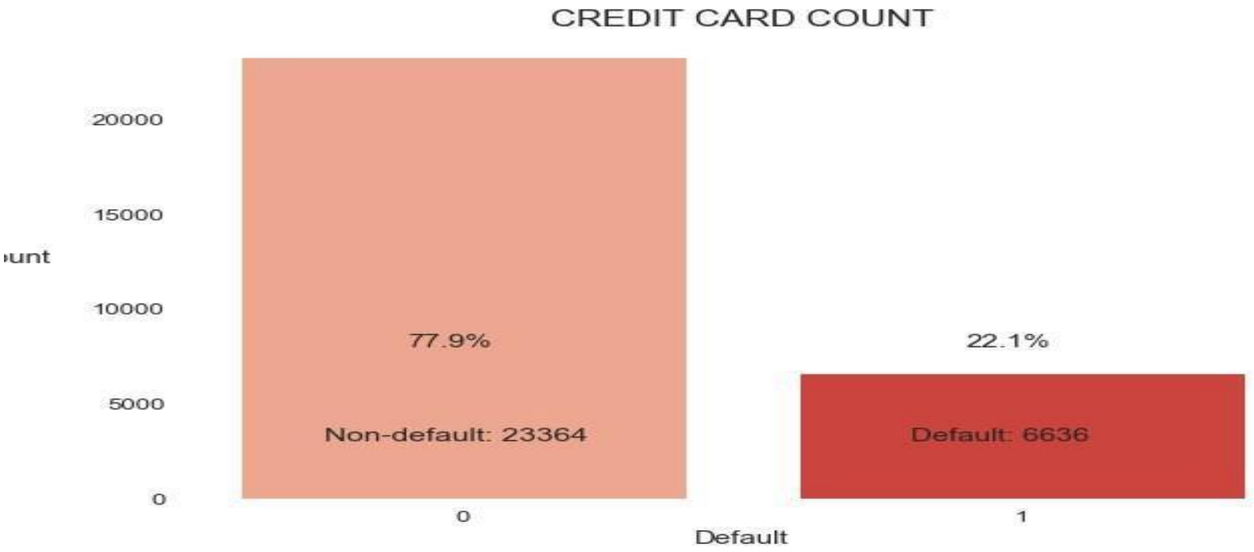
In case you're unclear on what defaulting on a credit card means, here's the gist: After you've failed to make a payment on your credit card for 180 days (or as decided by your credit card company), your issuer assumes you're probably never going to. At this point, the issuer can (and usually does) close your card, write off what you owe as bad debt and sell your account to a collections agency.

Project Objective

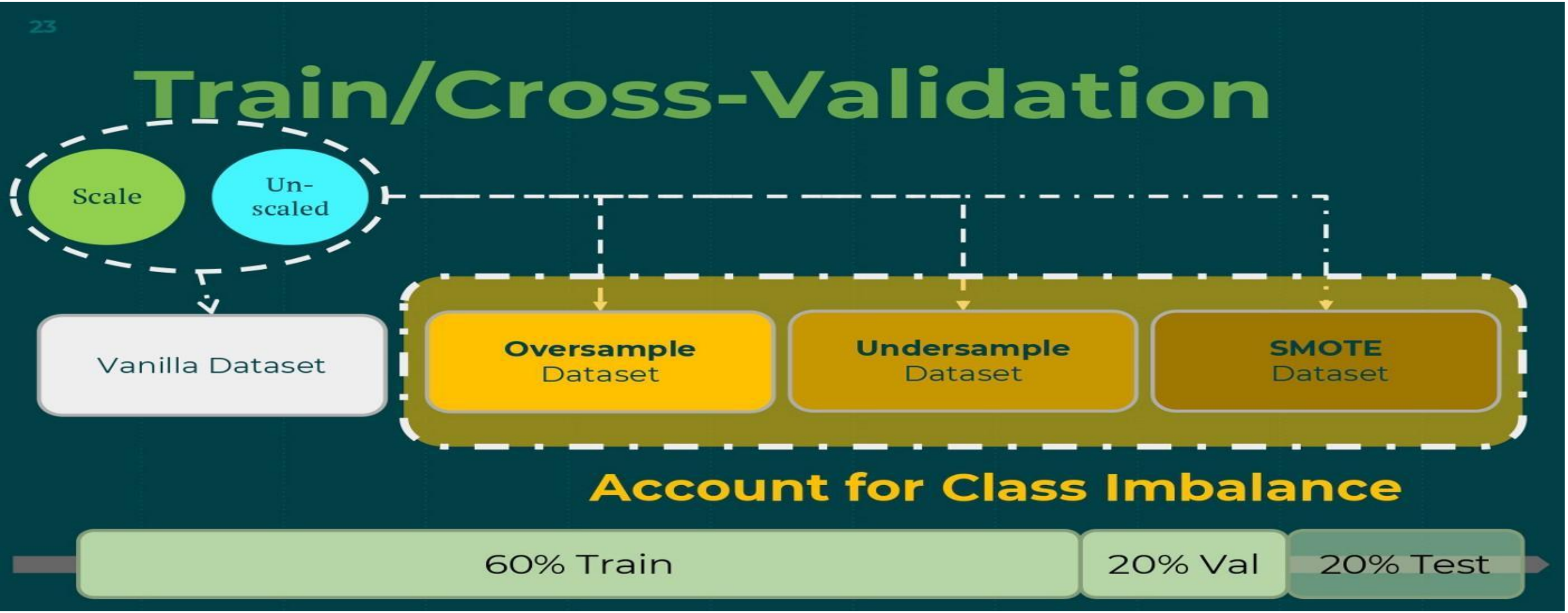
Therefore, this project aims to bridge this gap of uncertainty by utilizing a data-driven approach by using past data of credit card customers in ***conjunction with machine learning*** to predict whether or not a consumer will default on their credit cards.



Upon closer inspection of our dataset, we can see that there is some class imbalance, something of which we have to keep in mind when evaluating the efficacy of our model (i.e., we cannot use accuracy to measure our model's effectiveness). Additionally an imbalanced class will also affect the performance of the trained model. There are, however, a few ways to deal with class imbalance:



Each of the 5 models were trained on different variations of the dataset; scaled and unscaled oversampled, undersampled, SMOTE and vanilla datasets. As mentioned previously, there are a few ways to deal with imbalanced datasets. Oversampling, undersmapling and SMOTE were applied in this project.



23

Train/Cross-Validation

Scale

Un-scaled

Vanilla Dataset

Oversample Dataset

Undersample Dataset

SMOTE Dataset

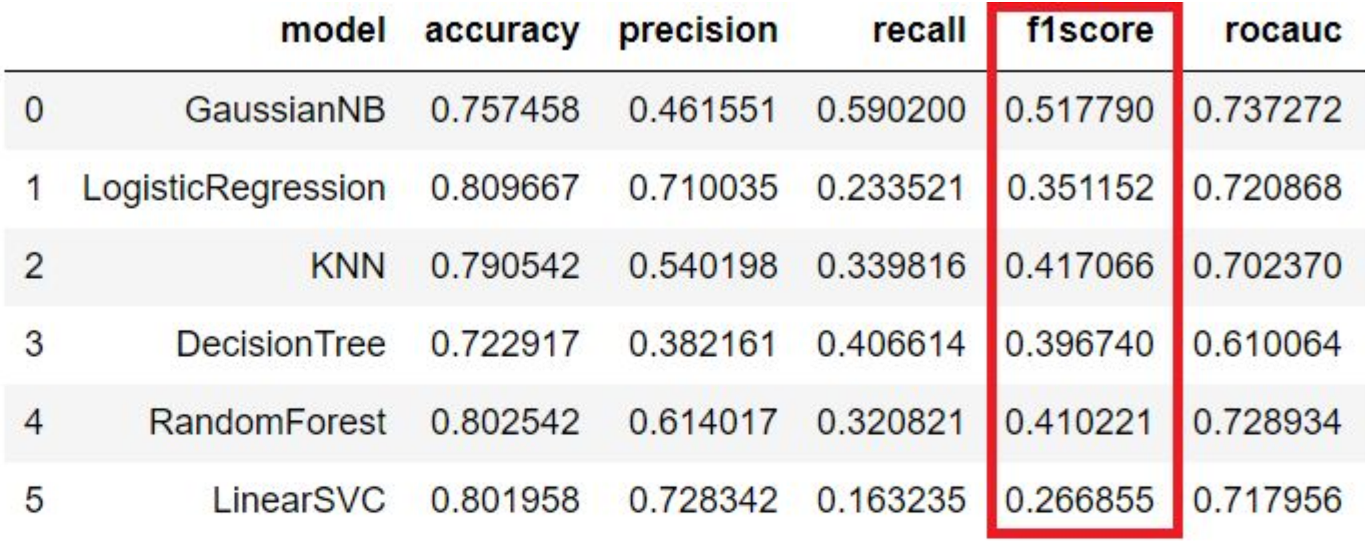
Account for Class Imbalance

60% Train

20% Val

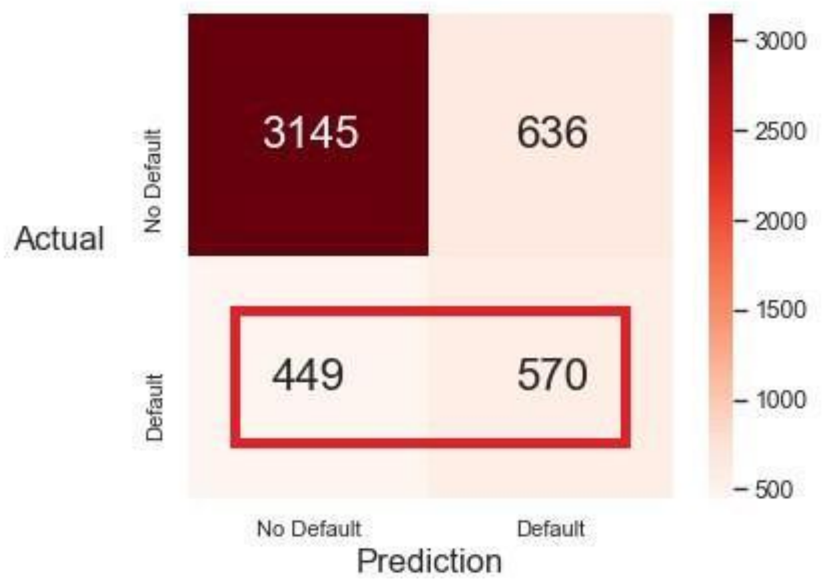
20% Test

After doing all the train/cv tests, the dataset that gave us the highest F1 score is the scaled vanilla dataset, with Gaussian Naive Bayes taking the lead at **0.518.** However, the models were all trained on their default settings. Therefore, I have chosen to further hyper-parameter tune the leading two highest F1 scoring models (**k-NN** at **0.417** and **RandomForest** at **0.41**) in a bid to obtain a higher F1 score than **0.518.**

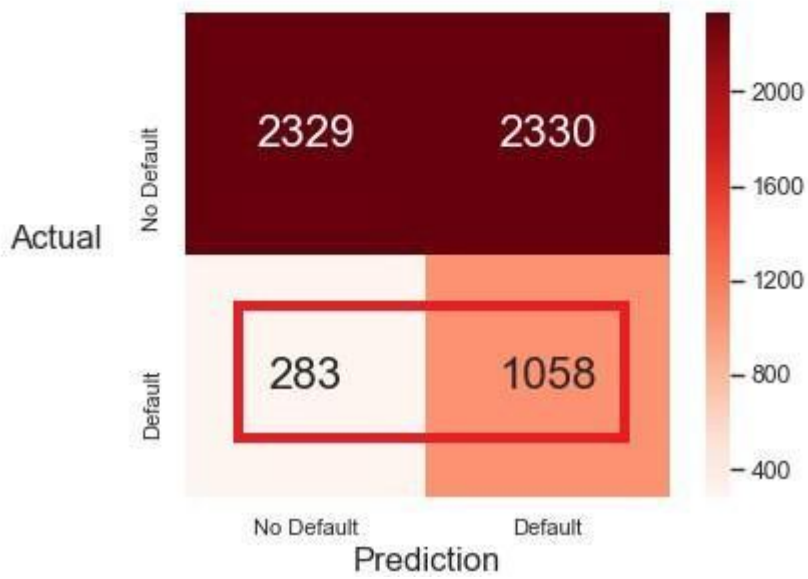


	model	accuracy	precision	recall	f1score	rocauc
0	GaussianNB	0.757458	0.461551	0.590200	0.517790	0.737272
1	LogisticRegression	0.809667	0.710035	0.233521	0.351152	0.720868
2	KNN	0.790542	0.540198	0.339816	0.417066	0.702370
3	DecisionTree	0.722917	0.382161	0.406614	0.396740	0.610064
4	RandomForest	0.802542	0.614017	0.320821	0.410221	0.728934
5	LinearSVC	0.801958	0.728342	0.163235	0.266855	0.717956

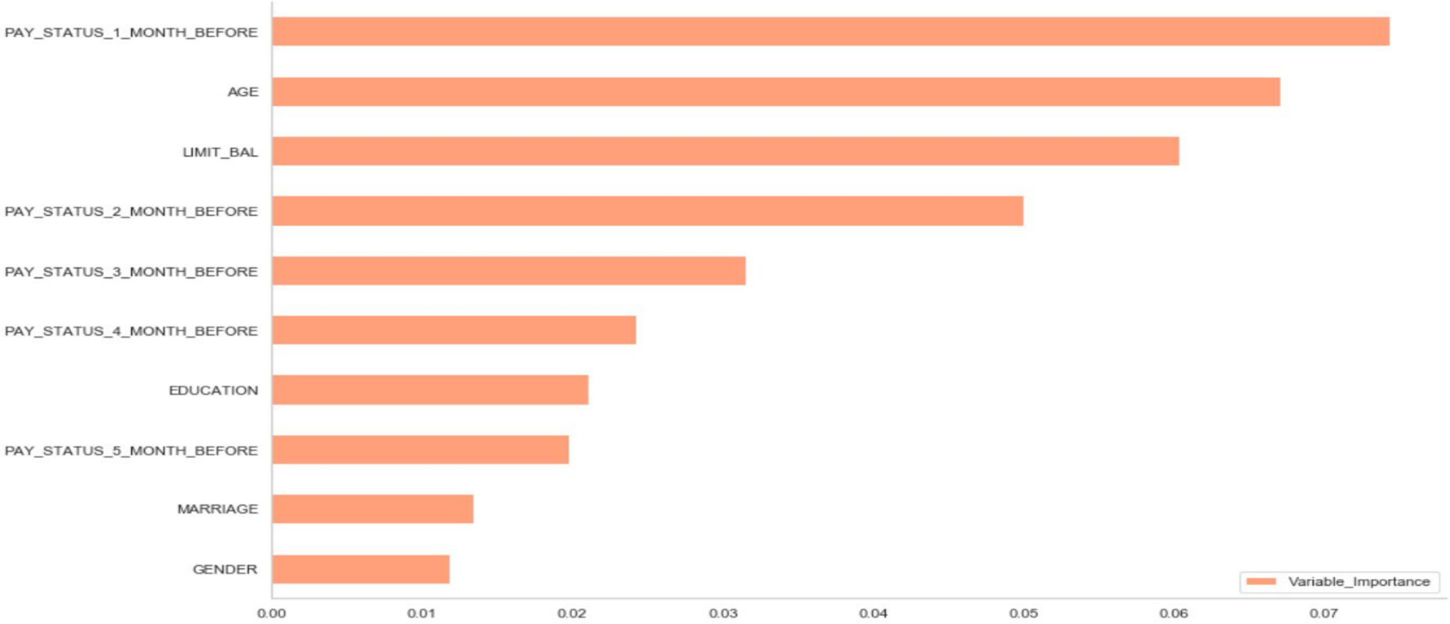
Upon plotting the confusion matrix of the Gaussian Naive Bayes with it’s default threshold of 0.5 (with F1 score of 0.517), we observe that it gives a very poor recall score. The above matrix tells us that the model only catches 56% of all the defaulters (570 out of 1019). Which means that ***misses 44% of all the defautlers***. This definitely does not spell good news for our hypothetical bank (very similar to credit card frauds) because we are letting a lot of non-loan-returning credit card defaulters go undetected. Therefore, there is a need to further optimize the threshold to increase the model’s recall.



Final tests show that our model actually performs even better on unseen data — giving us a higher ***recall*** score of **0.79** (1058 out of 1341). This means that our model is able to accurately catch around ***80%*** of all the defaulters.



From this graph, we can derive some interesting insights behind the behaviour of a defaulter. The top 3 telling drivers of whether someone is a defaulter or not boils down to their:



CONCLUSION

In this case, categorical columns like PAY_0 , …, PAY_6,MARRIAGE,EDUCATION may not have been represented the dataset in the best way. A better way could have been to do one-hot encoding (creating dummy variables) instead. Classifiers like RandomForests are great at segregating columns like these and could have resulted in a model that leads to better prediction of credit card defaulters.

AI

Capstone Project

TED TALK VIEWS PREDICTION

By – RAHUL SINGH WALDIA

TED

Well, I’ve analyzed a dataset of 2550 ted talks to get some answers for this question. I explored which of the available variables of a given talk, such as the number of comments, number of languages translated, duration of the talk,

number of tags, or day it was published online– are a strong predictor of its popularity, measured in number of views

After analyzing and regressing views over the other available variables in the dataset, some interesting associations were revealed.

What are features of a popular talk with high number of views?

High number of comments (naturally).

Translations in many languages (also, naturally).

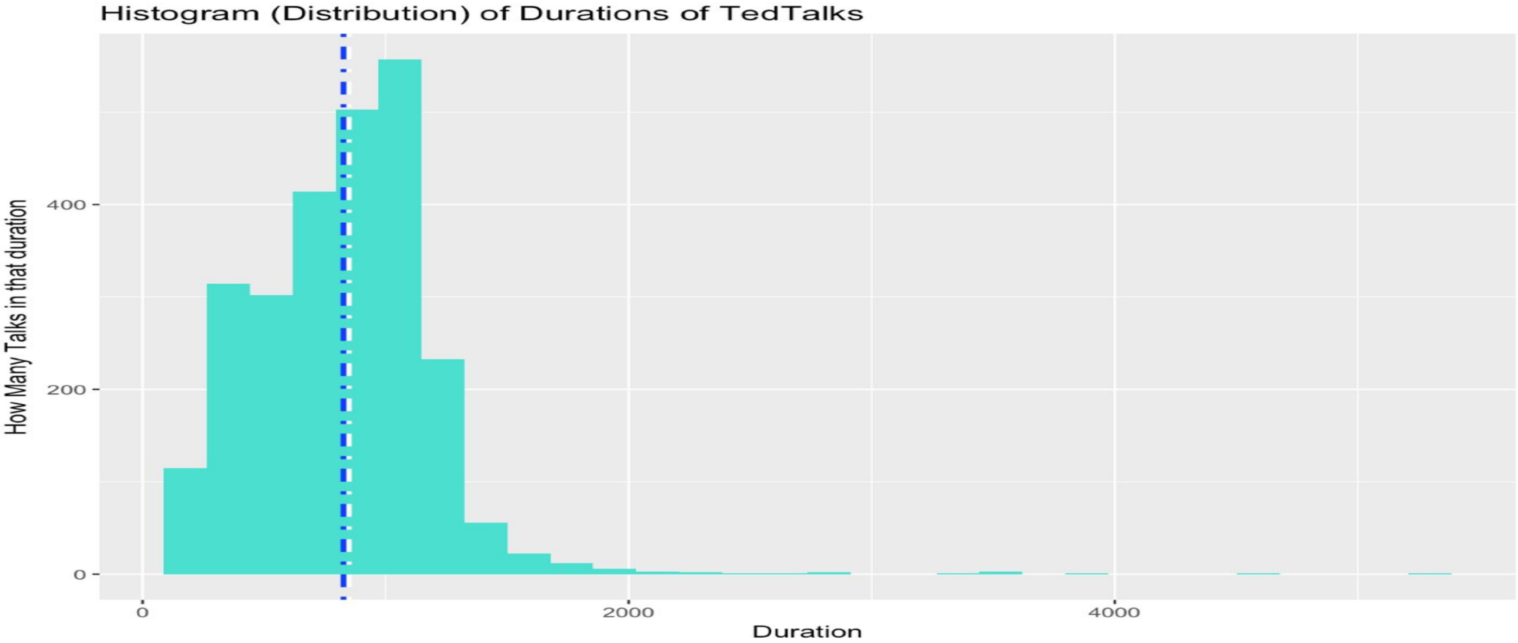
The combination of many comments *and* many translation languages yields a much higher view count than expected by each of them alone.

It *shouldn’t be too short*. Duration didn’t have a big effect at all, but if any, it was positive for longer talks. The most popular talks were between 8–18 minutes.

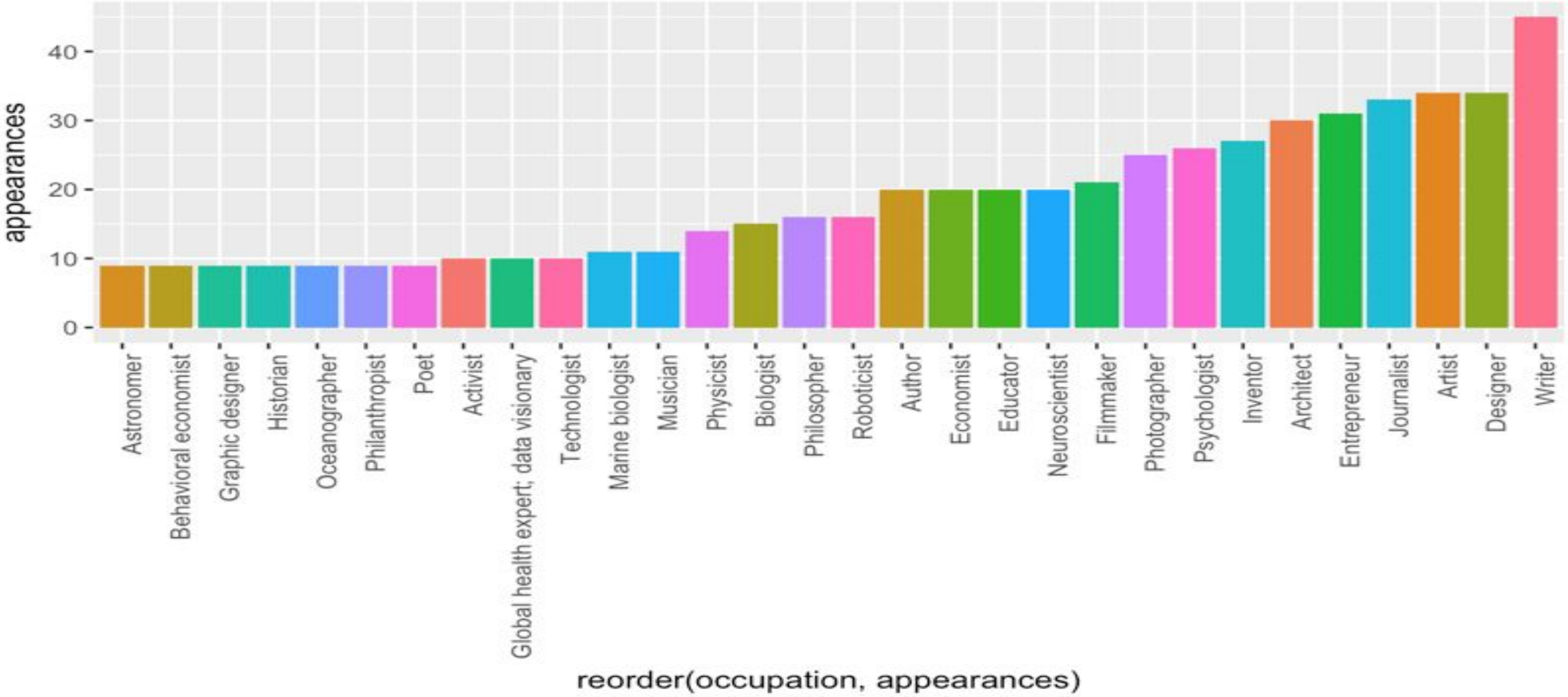
Higher number of tags, ideally between 3–8.

It would be uploaded on a weekday, preferably a Friday!

You may see some funky occupations yielding much higher than average views for their talks, such as: Neuroanatomist, Quiet revolutionary, Lie detector, Model, beatboxer, Vulnerability researcher, or Zen Priest. This isn’t representative, but they did yield the highest number of votes combined (which is an unfair game, but hey).

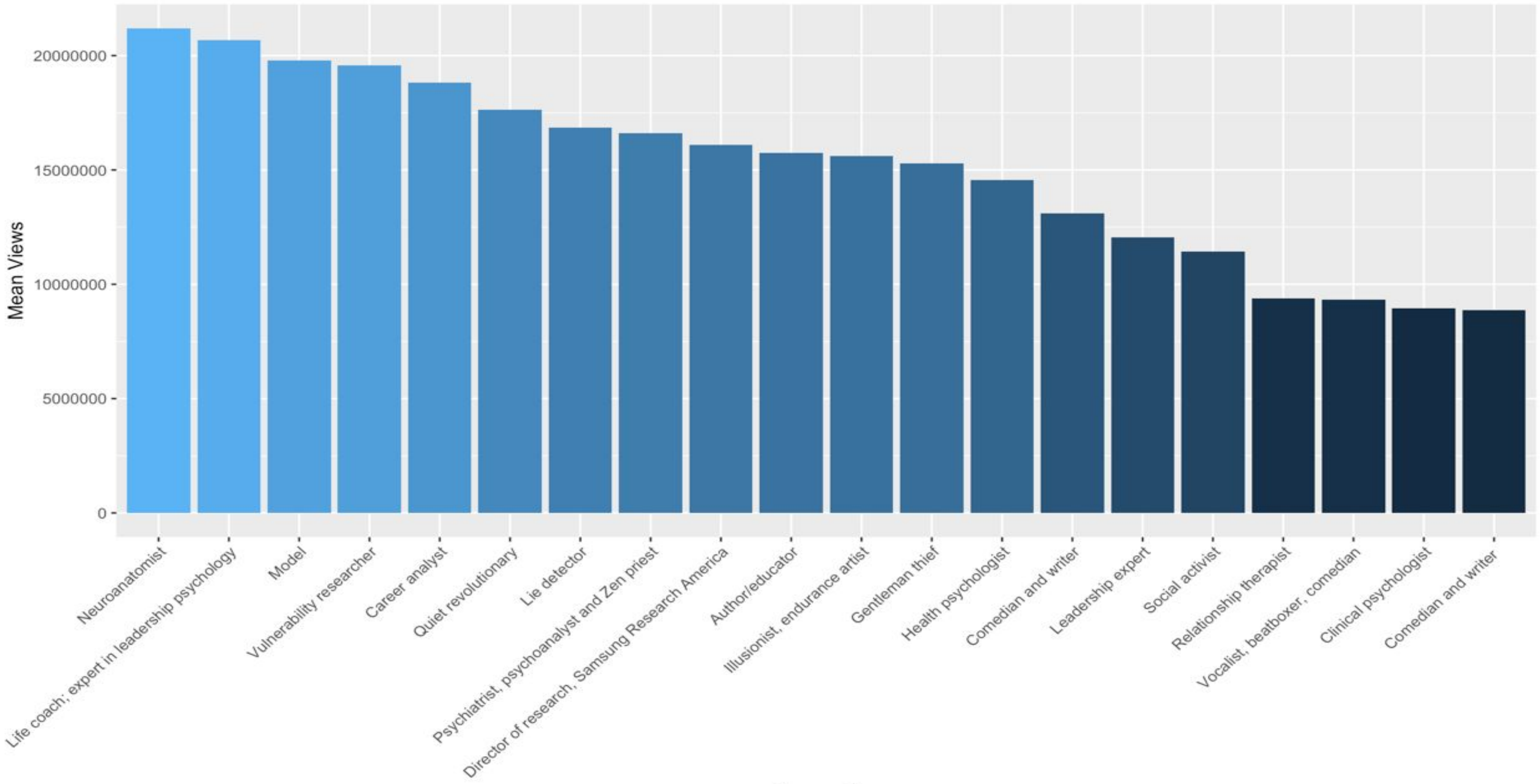


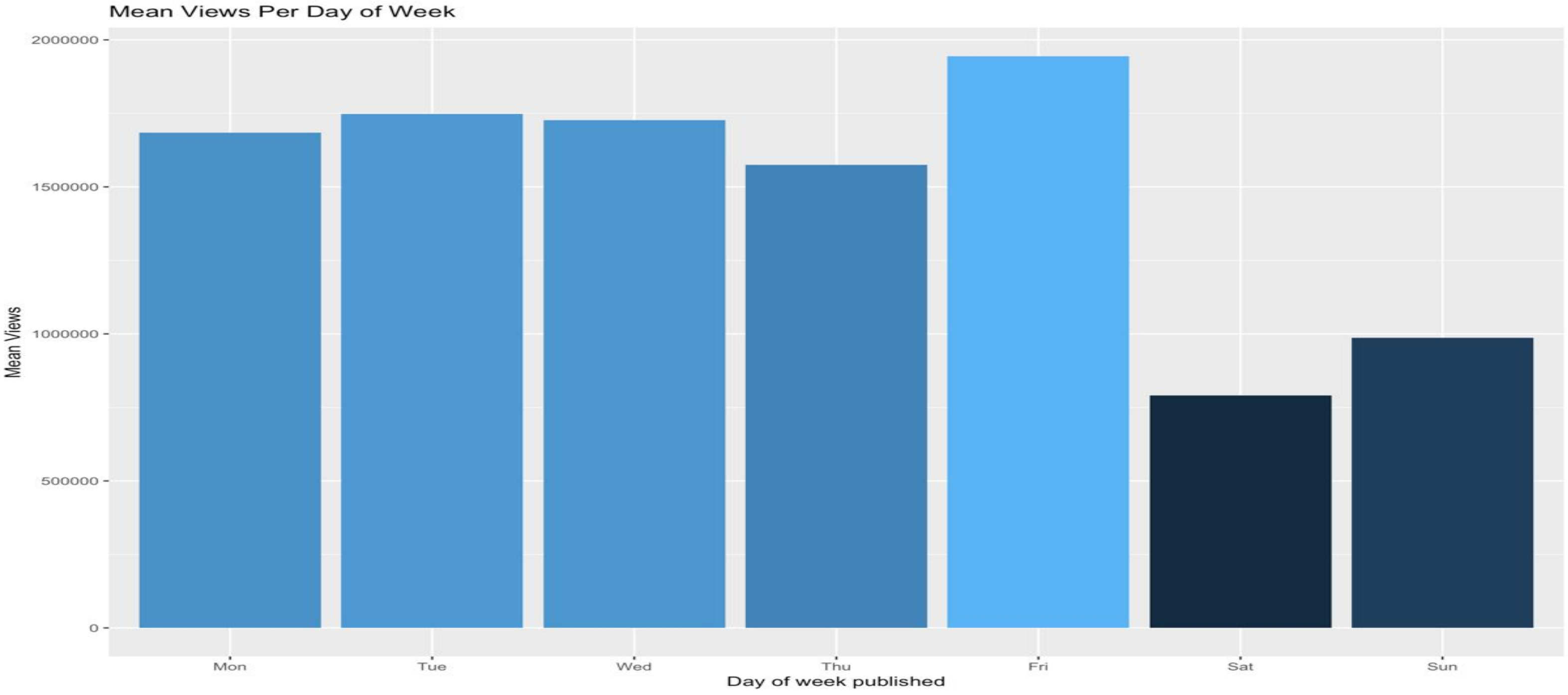
Duration of talks is closer to a normal distribution, but with a wide right-side tail of a few talks at longer duration, around a mean of 14 minutes and median of 12 minutes. Almost all talks range between 1–18 minutes (maximum length of a normal Ted talk).



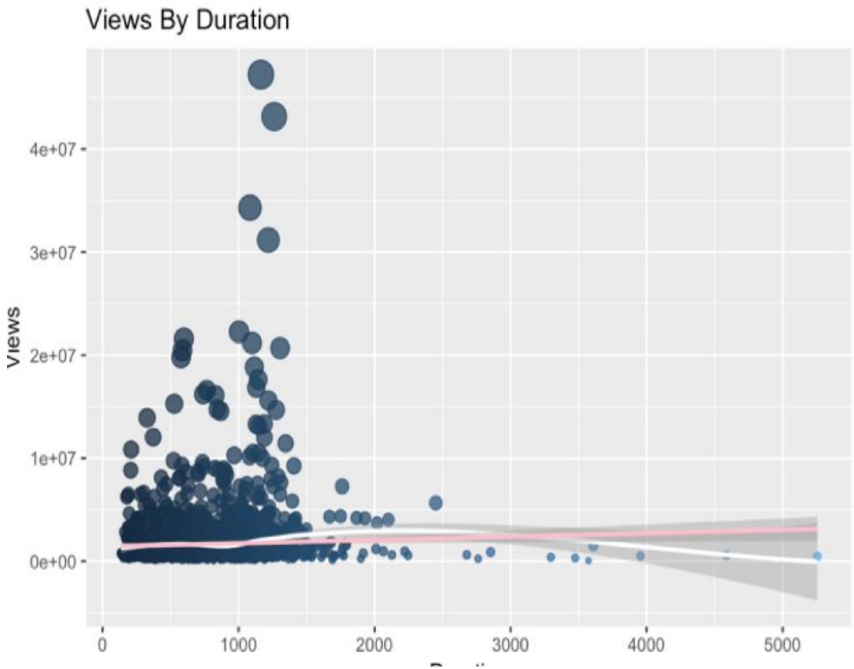
Apparently, writer is the most common occupation for a TED speaker! followed by other creative occupations and
number of speakers with that occupation:

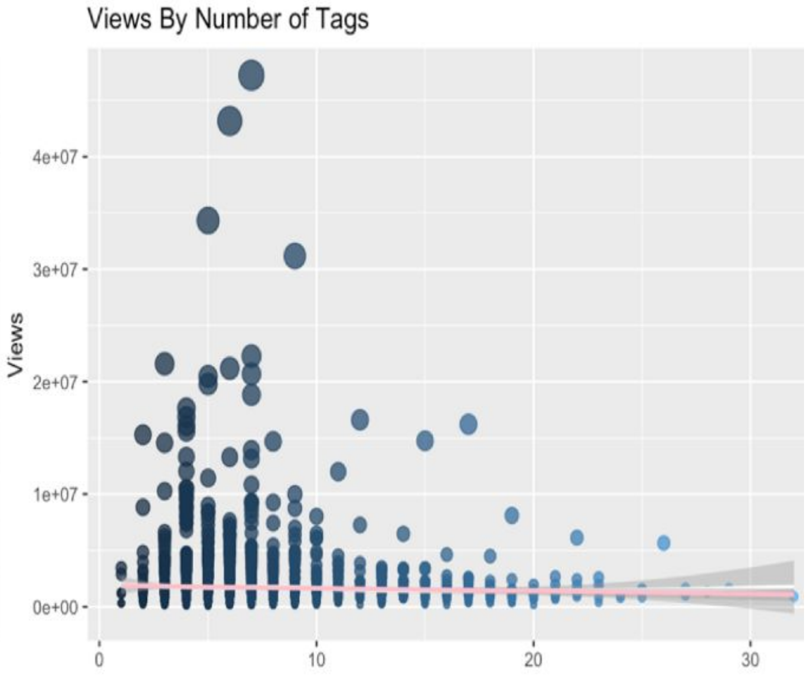
Occupation Vs Views





So, day of week does seem to have some association with average (mean) views! Ted Talks published on weekends seem to have much less views, with Saturday being the lowest, and Friday is the most popular day for ted talks published day.





Unsurprisingly, number of comments is, obviously, very well correlated with number of views, and so does number of languages — they all come from having many viewers. Thus, it is not “fair” to predict views based on these factors, and in the real world, we couldn’t use these parameters to predict, since they **are not causes** for more views, but they are also a result of many views, and a cause in a reinforcing feedback loop: the more comments, the more engaged the community is around the talk and likelier to spread; the more languages, the more viewers can watch; and the more viewers, the more audience there is to comment and translate. The rest had a small linear effect, where that didn’t deviate much, though small.

Conclusion

This limited model conveys correlation, not causation. It does not convey causal relationship well because the fundamental problem of causal inference is not well addressed with these variables, these predictors are *not independent from the y variable,* and they are highly correlated (mostly comments and number of languages which are the best predictors, naturally. I don’t believe that with these available numerical predictors we could have reached a causal inference. Next attempts might use the transcription of the talk to analyze the content, or audio to analyze the level of clapping, or the visuals in the talk and the clothing of the speaker to better predict using the content of the talk.