STOR 565 Final Project: Predicting Fake Job Postings

Leowell Bacudio
Michael Muller
Rui Wu

Problem of Interest

Most job postings today are made online. There have been cases of fake postings meant to collect personal information to sell and distribute to other companies. We would like to investigate online job postings and determine if it’s possible to predict if a posting is legitimate or fake.

Dataset Description

About

This dataset (from Kaggle) contains 18,000 job descriptions out of which about 800 are fake.

The data consists of both textual information and meta-information about the jobs.

Goal

Create classification models that uses text data features and meta-features to predict which job descriptions are fraudulent or real.

Important Variables

- title
- company_profile
- description
- requirements
- benefits
- location
- fraudulent (0 or 1)

Data Preparation

Reducing Dataset

Utilize only 4000 job postings from the dataset:

- Random sample of 3134 legitimate postings
- All 866 fraudulent postings

Transformations

NA values:

- Replace with empty strings

Text Descriptions:

- Strip all punctuations
- Change to lowercase characters

Extraction

Obtain length of text descriptions for certain variables. (5 new columns)

Obtain country of job posting based on location column. (1 new column)

Check if city for each job posting is one of top 5 cities in our sampled dataset, else "Other". (1 new column)

Natural Language Processing

Step 1

List of Words
TO
Numeric Array

Step 2

Numeric Array
TO
Single Numeric Value

NLP Step 1: Words to array of numbers

Goal

- Each text observation is a document.
- Each document contains many words.
- We need to assign a number to each word.

TF-IDF values

- **T**erm **F**requency: How many times a words appears in the document.
- **I**nverse **D**ocument **F**requency: 1 / proportion of documents that contain the word

Result

- Text turns into array of numbers.

NLP Step 2: Array to singular value

Goal

- Each observation is now array of numbers.
- Want to convert to a singular, numeric score.

Naive Bayes Classifier

- Make a classifier that uses the scores for each text observation to predict if text is of fraudulent class.
- NB Method allows access to raw percentage scores for each class.

Result

- Array of numbers for each observation turns into a percentage value.

Text

job_id	title	location	company_profile	description	requirements	benefits
6817	product in	DE, BE, Be	babbel enables any	we are looking f	requirementsy	we offer youpoter
16825	java tech l	US, CT, Ha	esolutions inc is a ta	titleÂ Â Â Â Â Â Â	requisition detail	skey responsibilit
7530	sheffield e	GB, , Sheff	established on the	under the natior	1618 year olds	career prospects
12534	it trainee	GR, I, Athe	urlc379aa631173e	who are wetrave	what are we loc	why travelplanet2
4278	graphic ar	US, DC, W	applied memetics ll	the graphic artis	the graphic artist shall be skilled in	
8311	english tea	US, TX, Da	we help teachers ge	jobs in china am	university degr	see job descriptio
7755	support w	GB, EDH, E	social care alba is th	social care alba i	key accountabil	this is your chance
14001	process er	US, CA, Ba	Â Â Â Â Â Â Â Â	we are a fullserv	experience preferred	pe registratic
13645	executive a	US, FL, Bo	marc bell capital pa	job descriptionÂ	must have exce	2014 employee be
708	bdc agent	US, TX, Dic	professional succes	as one of our liv	45 plus words per minute comput	
7351	growth ha	US, FL, tampa		this is an opport	skills excel and word excellent writ	
17705	admin assi	US, MI, Grand Rapids		job descriptionadministrative assistantdescriptiona		
741	health saf	US, CA, Ba	Â Â Â Â Â Â Â Â	health amp safet	duties and resp	what is offeredcor
8489	utc lead te	US, CA, Ba	jaco oil and refined	Â qualified candi	responsibilities	competitive comp
14319	developer	GB, LND, L	cloud 66 helps devs	cloud 66 is a techstars company building the best a		
1578	executive a	US, CA, Irv	happyfox is a young	happyfox is a sta	be absolutely m	competitive payÂ
17811	business d	US, ,		we have the demand we are looking for people tha		
11387	senior dev	DK, 84, KÃ	at founders we crea	be part of buildi	need to haveha	the adventure we
12940	applicatior	US, CA, Re	our passion for imp	the company es	the ideal candid	our culture is anyt
17658	data entry	US, CA, LOS ANGELES		immediate open	some clerical ar	vacations holiday
13103	parttime w	AU, NSW, Sydney		parttime work from your place		flexible schedule 65
5585	structural	US, TX, Houston		why choose aec	minimum requirements	qualificat
1383	van forem	US, IL, Eas	federal has been in	driver career op	job requiremen	all drivers receive
15137	executive a	GR, I, Athe	optimal business ac	on behalf of our	excellent verbal and written comm	
15252	print desig	US, PA,	printfreshÂ is a lead	the ideal applicant will have 35 years of experience		
1190	carerseni	GB, NYK, H	inception recruitme	must have previ	the ideal candidate will have a stro	

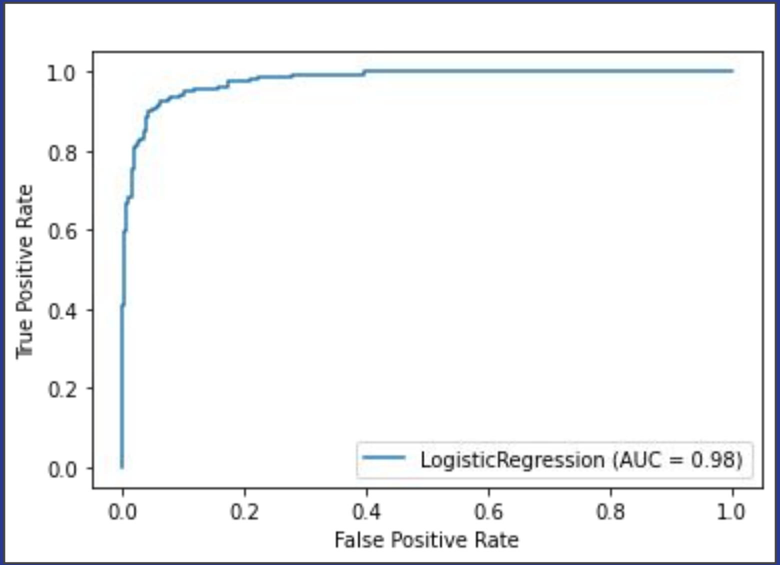
Likelihood Text is Fraudulent

likely_title_fraud	likely_profile_fraud	likely_desc_fraud	likely_req_fraud	likely_ben_fraud
0.003669685	1.16E-48	9.11E-27	7.41E-15	2.11E-14
0.006502364	4.29E-10	2.22E-13	2.65E-22	0.000319338
6.93E-09	1.03E-84	1.91E-21	2.86E-10	1.14E-05
0.001046579	4.10E-08	1.63E-19	9.14E-17	8.72E-12
0.058667216	1.29E-89	1.06E-13	8.32E-10	0.000319338
7.95E-07	0.999952257	6.32E-22	4.94E-16	2.43E-05
0.012579502	2.11E-42	3.16E-21	1.73E-29	1.06E-16
0.348779284	1	1	0.999999203	0.000283497
0.025376563	3.69E-13	1.42E-10	1.23E-05	2.97E-05
0.065540173	5.73E-12	5.30E-15	3.92E-12	0.000283497
0.005054239	1	2.28E-27	9.80E-17	0.000369926
0.124901442	1	0.999999984	0.000231041	0.000369926
0.246587435	1	0.999984586	1	1
0.631574045	1	0.62362458	1	1
0.007588341	0.277880008	7.28E-39	0.000231041	0.000369926
0.024655447	1.64E-17	2.74E-19	9.95E-09	6.14E-05
0.072274837	1	1	0.000187605	0.000279675
0.001999583	7.60E-16	1.27E-25	3.08E-18	7.21E-10
0.001262796	3.90E-28	1.83E-26	2.63E-11	1.78E-28
0.991379653	1	0.859114837	1.67E-10	0.000556078
0.955950358	1	7.89E-11	0.000199722	0.000285007
0.302630912	1	0.004219086	3.37E-06	0.000285007
0.070815315	8.95E-11	1.18E-16	1.13E-06	0.009372286
0.025062998	8.53E-83	8.17E-18	2.81E-05	0.000285007
0.002947776	7.89E-14	8.15E-17	0.000199722	0.000285007
0.285959929	0.001883282	7.06E-17	5.28E-08	0.000328407

Modeling

Model 1

Logistic Regression



Most Impactful Statistically Significant Variables

Not Fraudulent

- Bachelor’s Requirement
- Management Jobs
- Executive experience requirement
- Customer Service Jobs

Fraudulent

- Benefits fraud %
- Requirements fraud %
- Description fraud %
- Accounting / Auditing Jobs

Summary

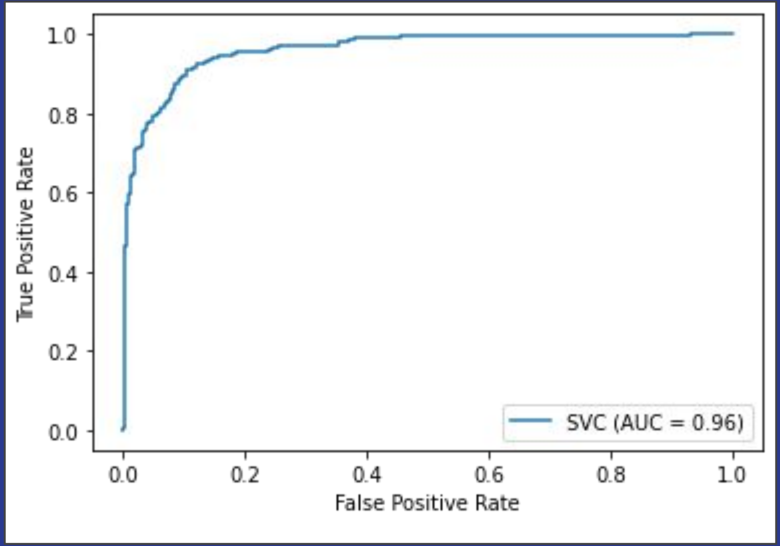
First model built to test for simplicity.

Full model: 91.8% test accuracy.

Feature selected model: 91.8% test accuracy.

Model 2

SVM

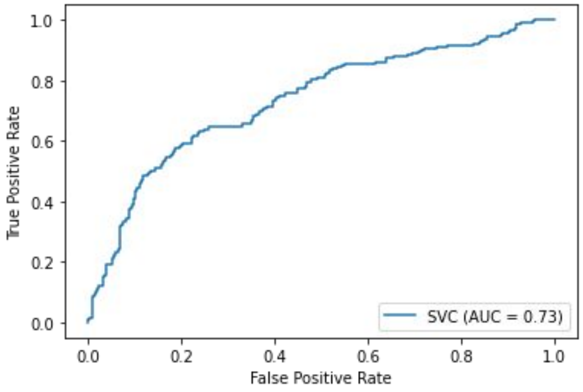


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
SVC (AUC = 0.96)

Rbf Kernel

Test Accuracy: 76.5%

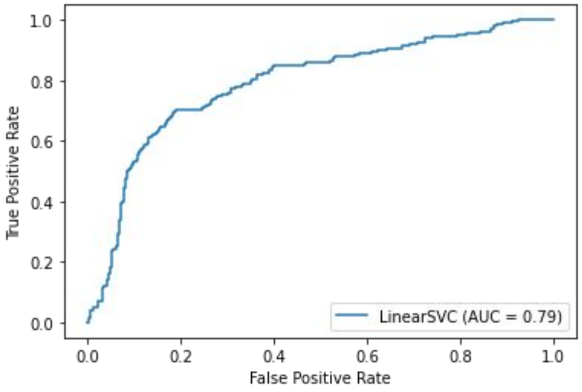
SVC with all features = not good



Linear Kernel

Test Accuracy: 70.6%

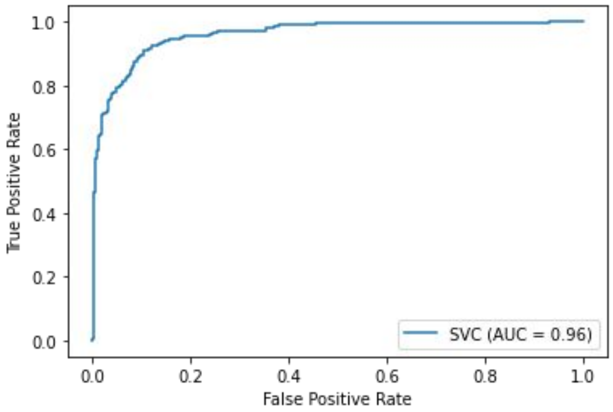
Just guessing non-fraudulent would get higher than 80%!



Rbf Kernel

Reduced to 20 features.

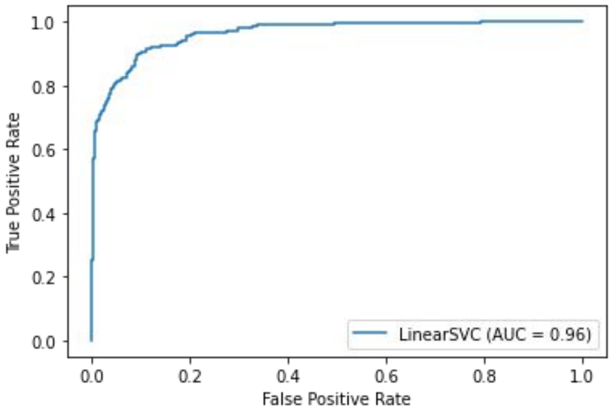
Test Accuracy: 91.1%



Linear Kernel

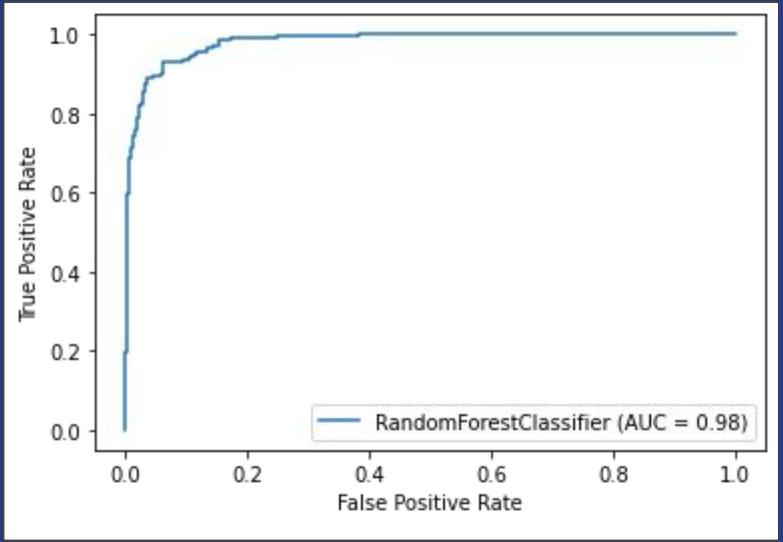
Reduced to 20 features.

Test Accuracy: 91.8%



Model 3

Random Forest



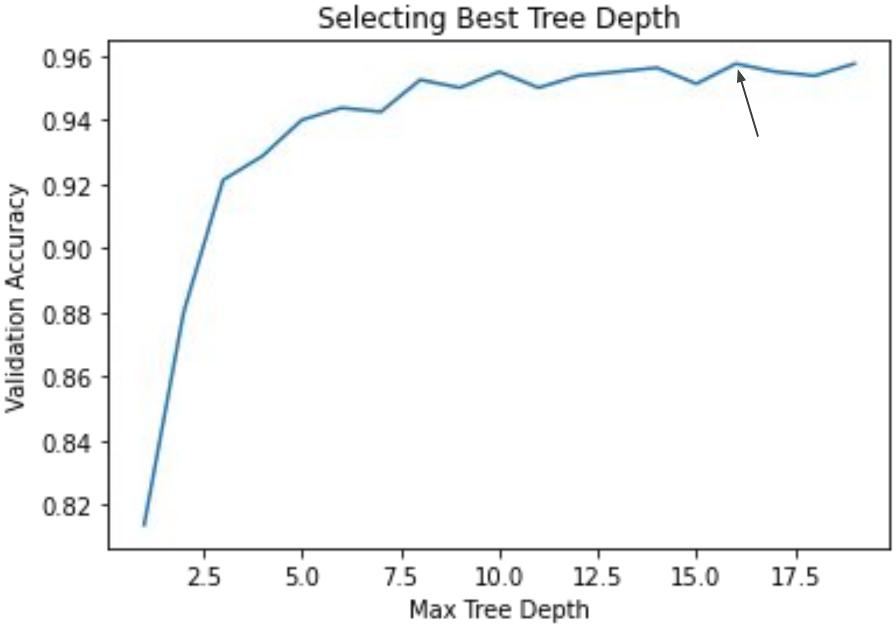
RESULTS

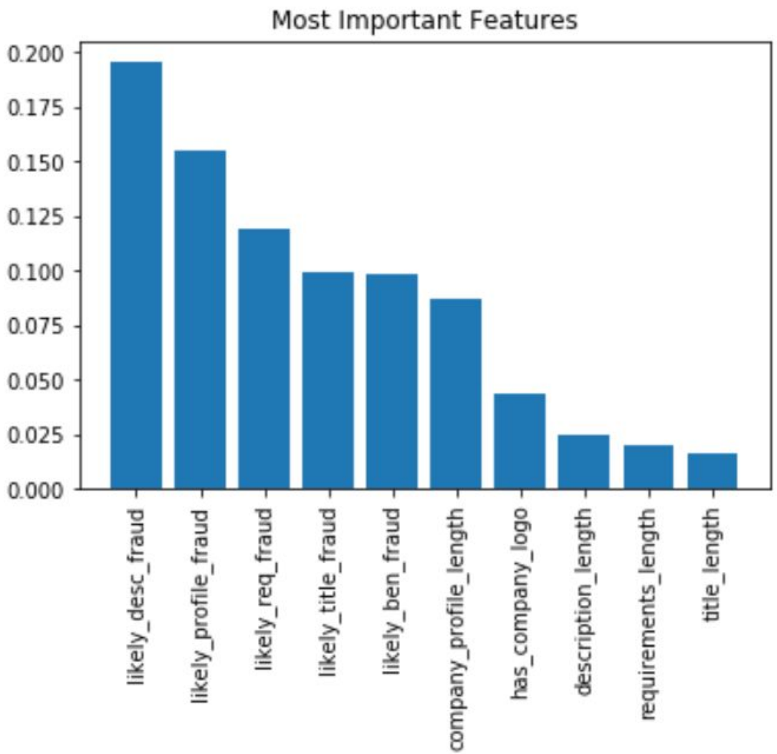
Best tree depth = 16

Validation accuracy: 95.0%

Test accuracy: 93.6%

BEST MODEL!





ANALYSIS

NLP related columns had most weight as expected

Description and profile had more weight than requirements, title or benefits.

Notable predictor: Has company logo

Conclusion

Most Important Predictors

NLP Columns	Non-NLP Columns
Description + Company profile very important • Unique to each job • Hard to fake **Requirements + Title + Benefits not as important** • Requirements are often similar • Title has few words • Benefits not commonly included in dataset	**No company logo? Probably fake** **Required education** • If listed, then probably in the clear **Accounting and auditing jobs more likely to be fraudulent!**

Questions?