Predicting Campaign Outcomes

ML Modeling Approaches to Predict Successful Kickstarters

CASE STUDY

APRIL 22, 2022 | KATHLEEN TRINH

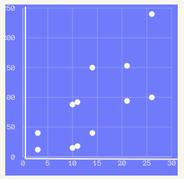
THE ASK

WITH A GIVEN DATASET, USE PREDICTIVE MODELING TO DESIGNATE KICKSTARTER CAMPAIGNS AS ULTIMATELY DESTINED FOR SUCCESS OR FAILURE.

ADDITIONALLY, ENGINEER FEATURES FROM THE DATA, PARTICULARLY THE NAMES OF THE CAMPAIGNS, TO INCREASE THE ACCURACY OF THE MODELING APPROACH BEYOND THE SCOPE OF THE ORIGINAL VARIABLES PROVIDED.

AGENDA

Solution



Data Preprocessing, EDA, & Baseline Model Performance

Feature Selection & Transformation Methods

NLP Features

Final Model Performance

Conclusion

Project We Love

Building Community One Taco at a Time

Help a local taco truck expand and reach new heights

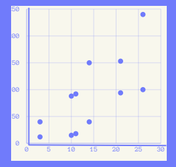
by Derrick Braziel

$15,957 pledged
79% funded
17 days to go

Food Trucks Cincinnati, OH

SOLUTION TACTICS

- Leverage NLP derived features to create variables that will help to increase the performance metrics of baseline modeling approaches (Logistic Regression & Random Forest).
- Additionally, account for multicollinearity within original dataset and perform dimension reduction via filtering based off of Pearson correlation.

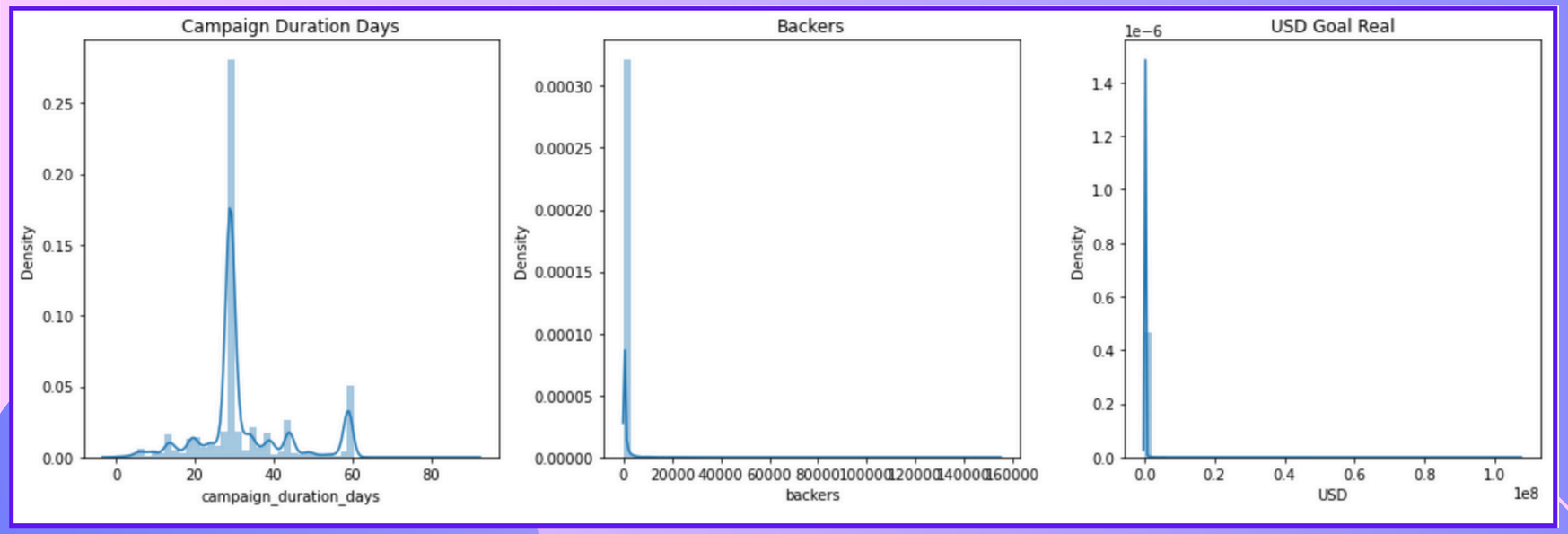


Initial Preprocessing

- Created new variable from launch and end dates of campaign
- Performed binning on respective variables' distributions
- Removed variables with potential for information leakage (e.g. pledged vs. goal $ amounts).

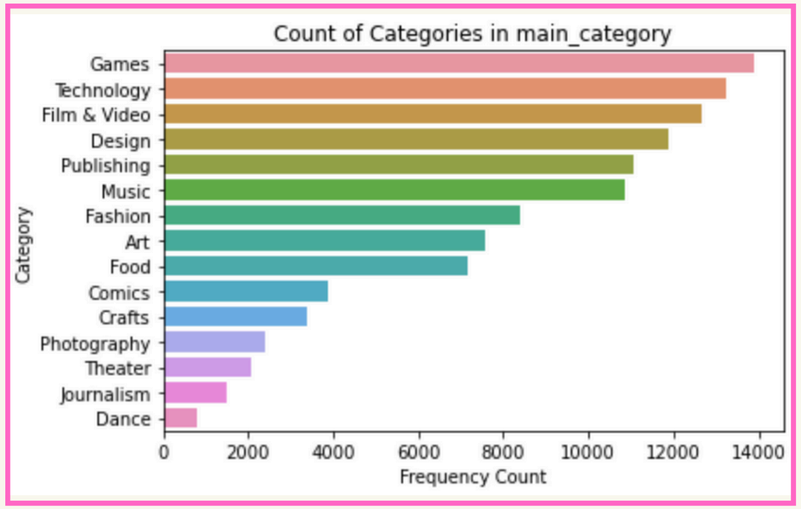
campaign_duration_days	campaign_duration_days_0_30	campaign_duration_days_30_74
24	1	0
20	1	0
52	0	1

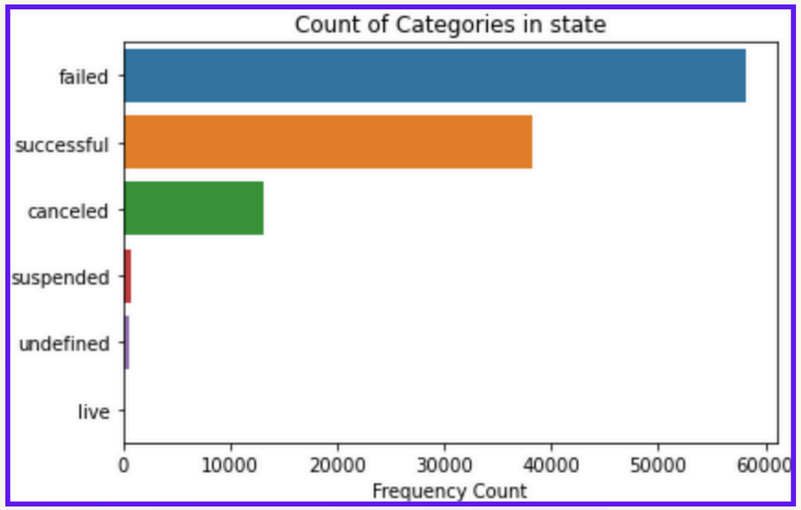
backers	backers_0_20000	backers_20000_40000	backers_40000_60000	backers_60000_80000	backers_80000_100000	backers_100000_120000	backers_120000_140000	backers_140000_160000
36781	0	1	0	0	0	0	0	0
63758	0	0	0	1	0	0	0	0
20787	0	1	0	0	0	0	0	0

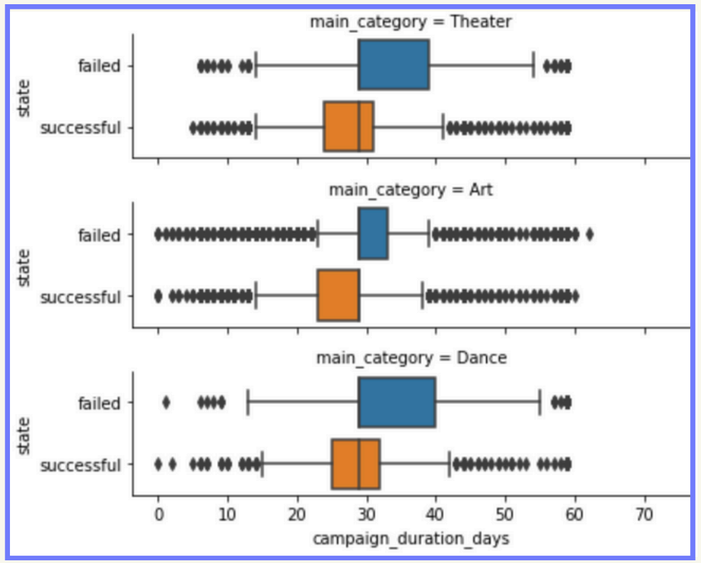


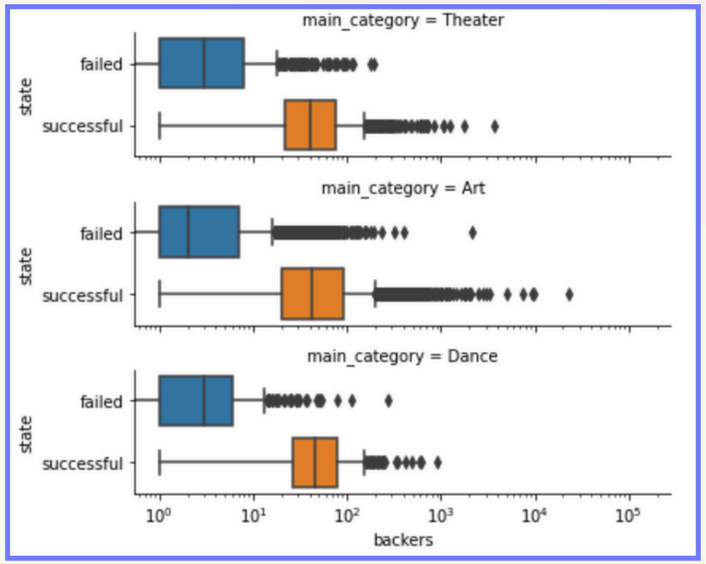
EXPLORATORY DATA ANALYSIS

Discovering insights, such as correlated variables, to help create an optimized dataset for our model.







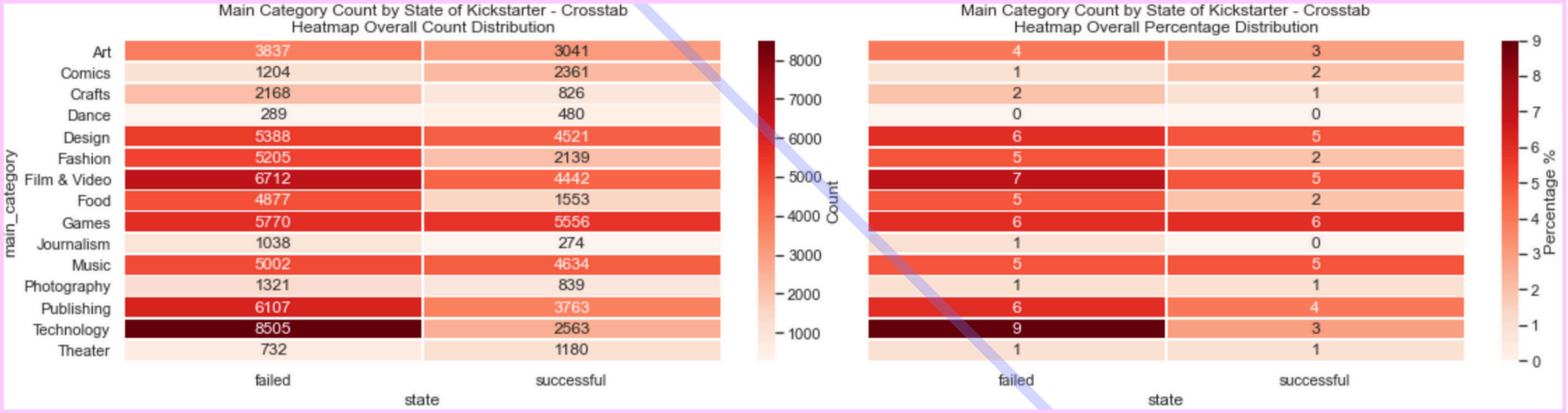


SOME OVERARCHING TRENDS:

- Most campaign durations are within 30-35 days.
- Most campaigns' pledged amounts tend to not meet goals, which is also indicated by the majority of campaigns failing. These campaigns often have a low number of backers (skewed to <1K).

SUCCESS

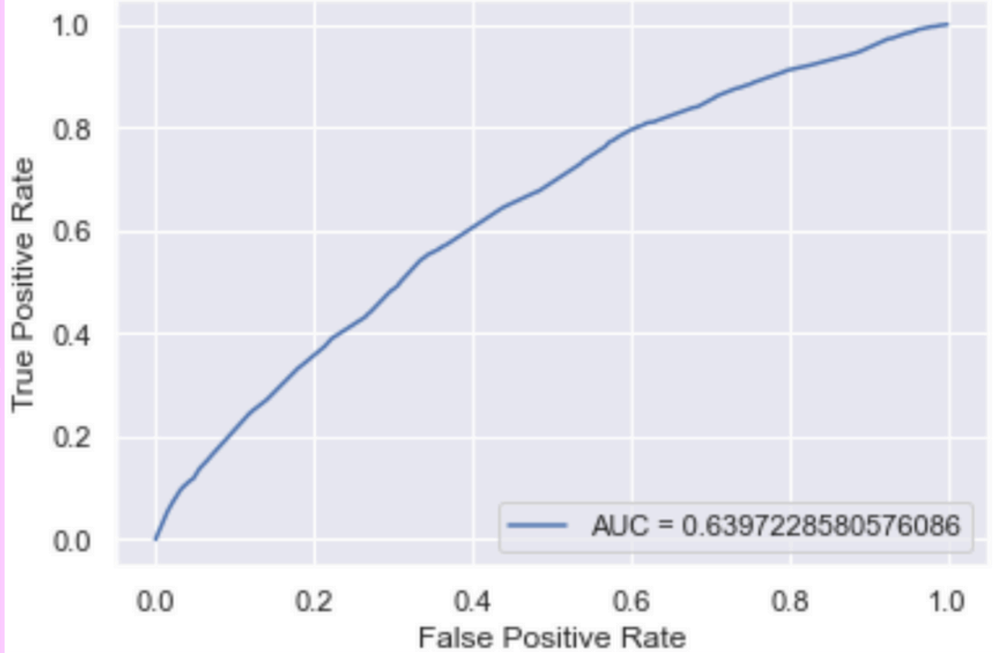
The "Games" category had the highest ratio of successful campaigns.



FAILURE

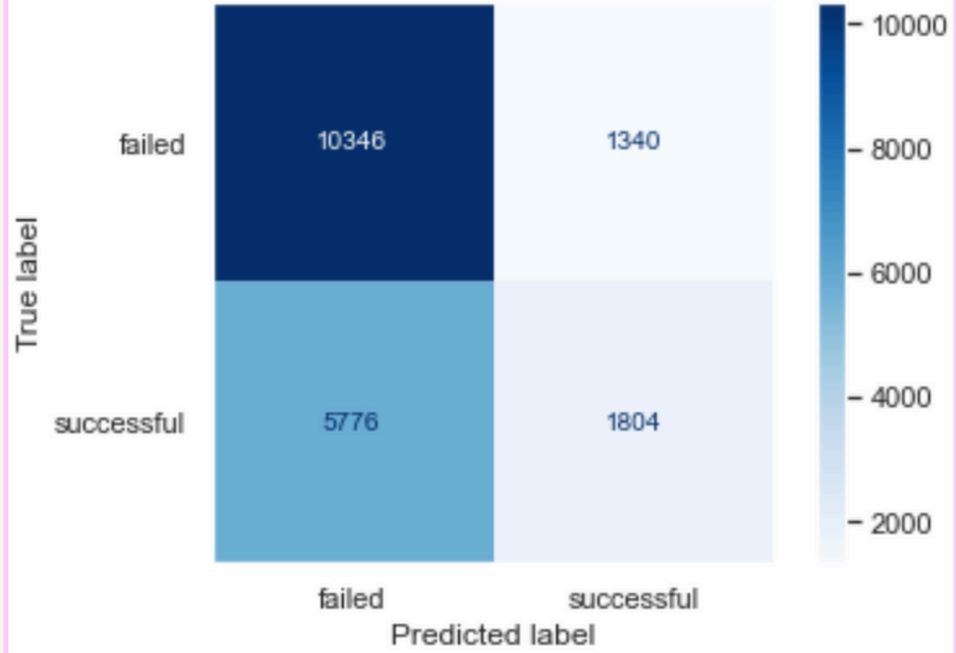
Conversely, the least successful campaign type is "Journalism".

INITIAL LOGISTIC REGRESSION MODELING PERFORMANCE METRICS



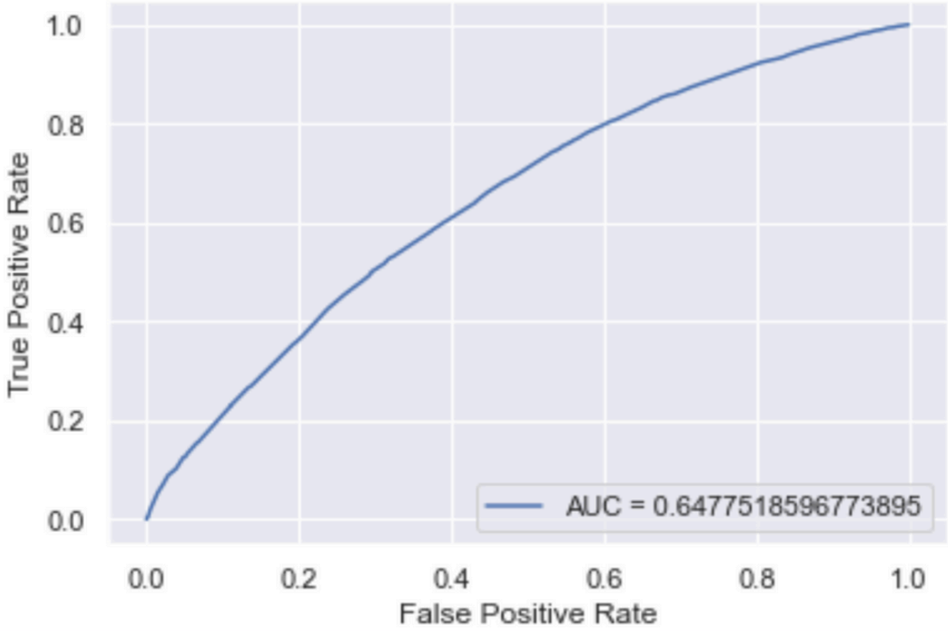
AUC

	precision	recall	f1-score	support
0	0.89	0.64	0.74	16122
1	0.24	0.57	0.34	3144
accuracy			0.63	19266
macro avg	0.56	0.61	0.54	19266
weighted avg	0.78	0.63	0.68	19266



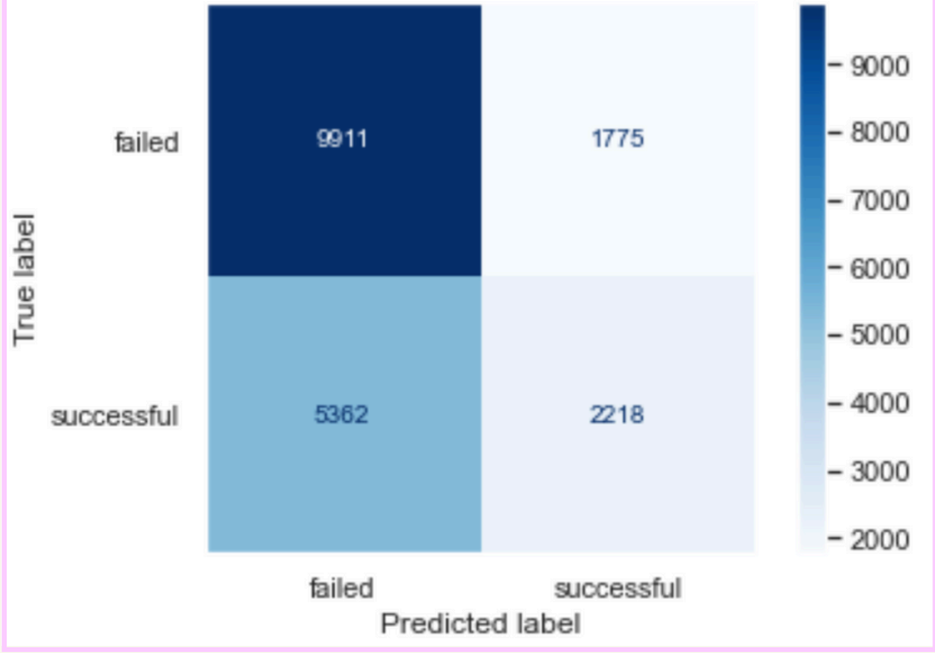
CONFUSION MATRIX

INITIAL RANDOM FOREST MODELING PERFORMANCE METRICS



AUC

	precision	recall	f1-score	support
0	0.85	0.65	0.74	15273
1	0.29	0.56	0.38	3993
accuracy			0.63	19266
macro avg	0.57	0.60	0.56	19266
weighted avg	0.73	0.63	0.66	19266

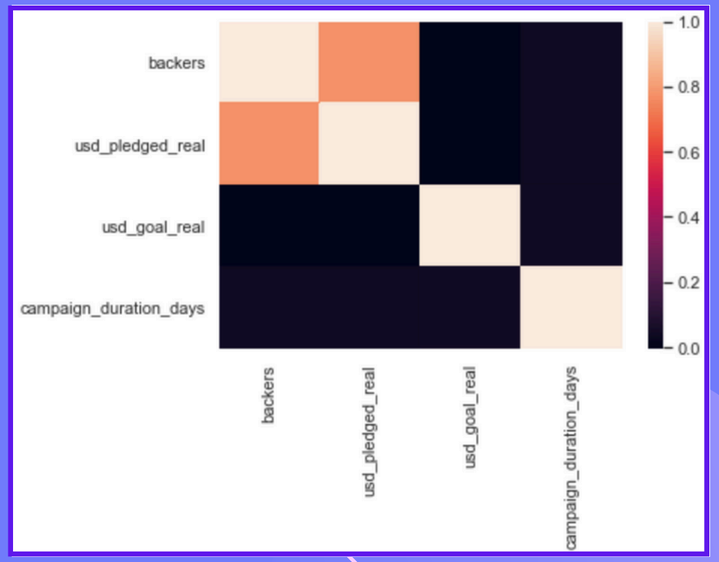


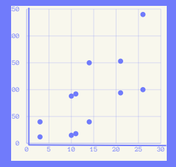
CONFUSION MATRIX

DISCOVERING MULTICOLLINEARITY & ADDRESSING VIA DIMENSION REDUCTION

DIMENSION REDUCTION METHOD: FILTERING

- Use correlation threshold of >0.5 to detect variables to exclude
- Caveat that we select features that do have high correlation with target variable, but only if they do not count as information leakage (e.g. not using campaign $ amounts as it erroneously inflates model performance)



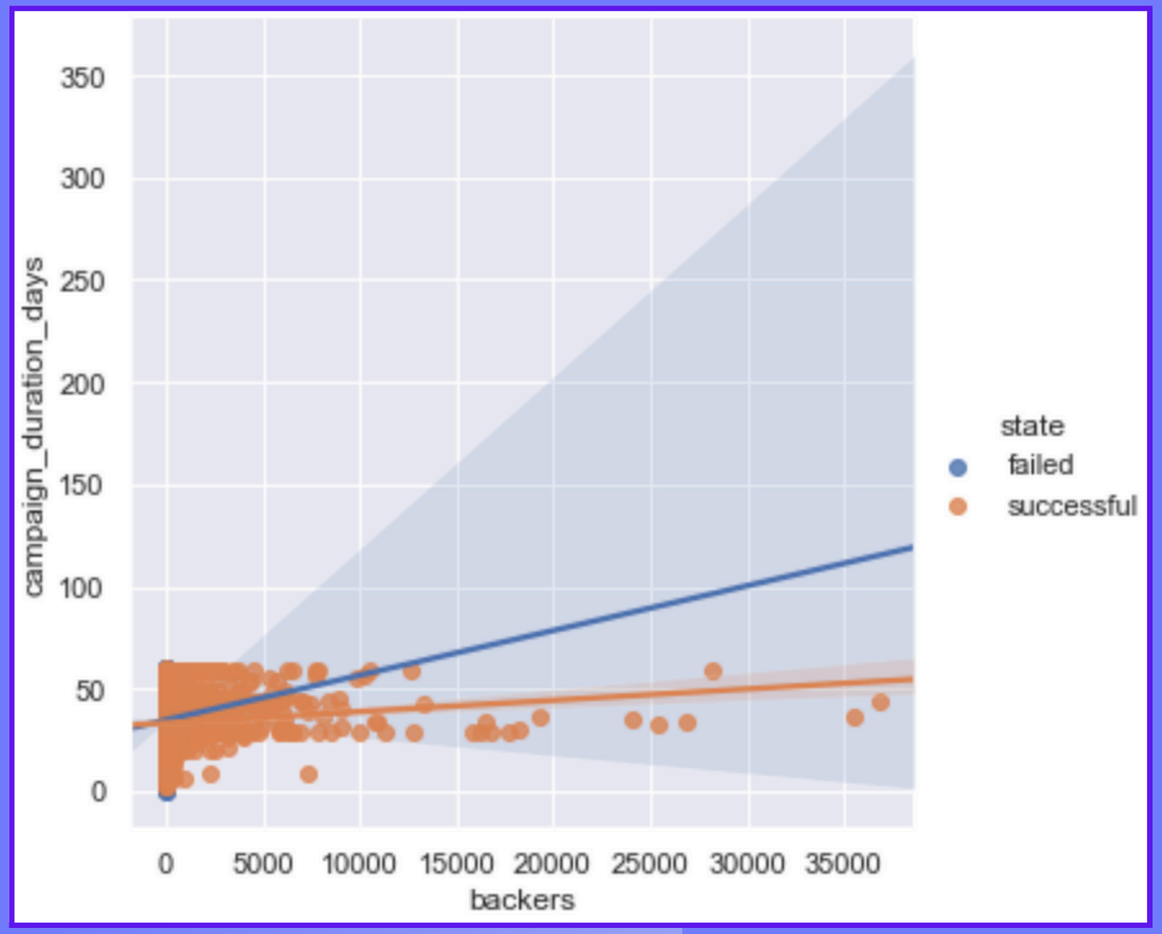


EXAMPLE OF PEARSON CORRELATION PLOT

Used to determine features to exclude as a means to select relevant, optimal features from the given dataset.

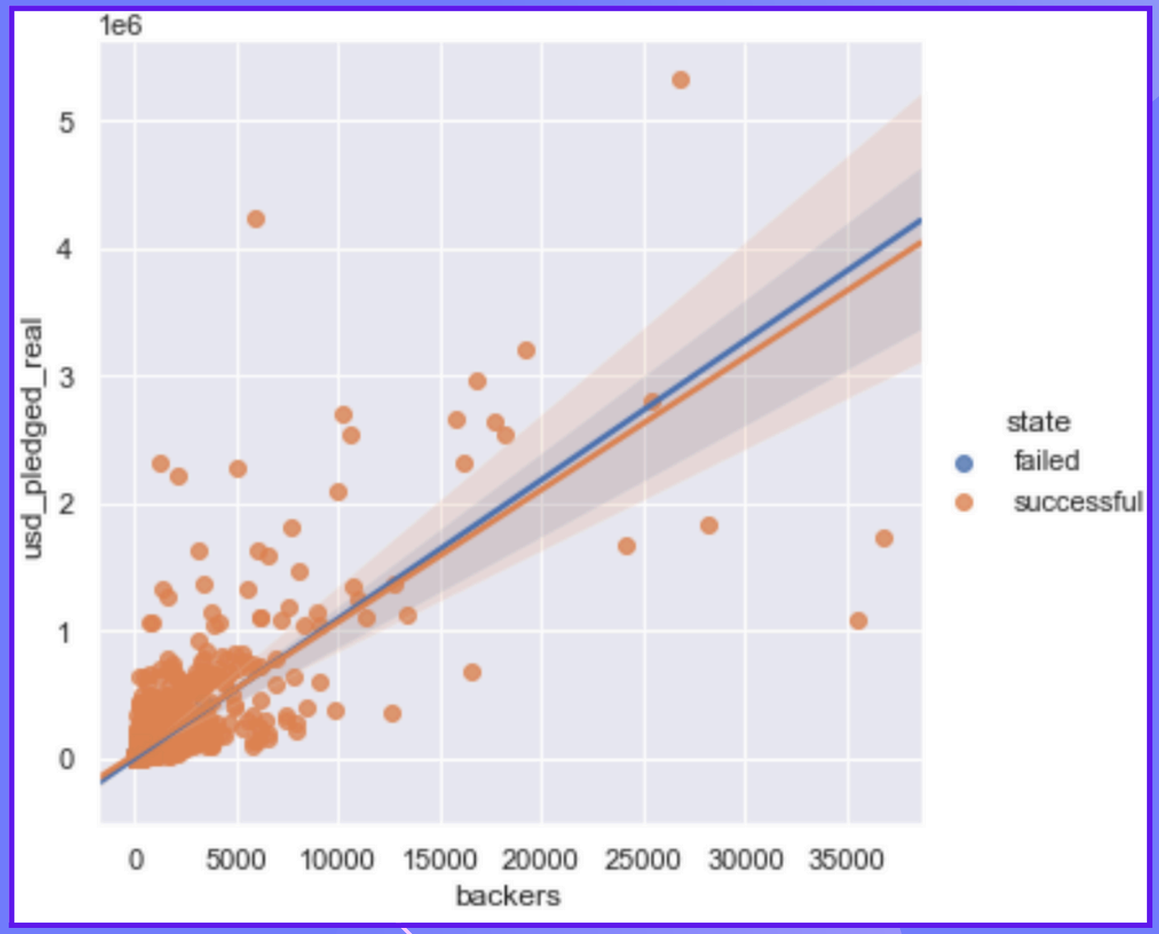
```
# Calculate Pearson Correlation Coefficient and p-value between 'backers' and 'campaign_duration_days'
stats.pearsonr(kickstarter_subset['backers'], kickstarter_subset['campaign_duration_days'])
```

```
(0.03489162026182735, 0.00024116798094624215)
```




```
# Calculate Pearson Correlation Coefficient and p-value between 'backers' and 'usd_pledged_real'
stats.pearsonr(kickstarter_subset['backers'], kickstarter_subset['usd_pledged_real'])
```

```
(0.7757848580660223, 0.0)
```



RATIONALE OF FEATURE TRANSFORMATIONS

Transformation of categorical values into numeric labels by applying ordinal encoding with a hard-coded dictionary was ultimately selected over one-hot encoding or regressed continuous values due to the following rationale:

- One-hot encoding leads to the formation of many new columns, yielding a wide and sparse dataset, which leads to difficulty for classification models to assess variables in high-dimensional space along with potential memory fallout issues.
- Regressed continuous values ideally work on features without multimodal distributions (more than one local maximum), which, unfortunately, many of our features had.

MemoryError

```
logistic_regression_model1 = LogisticRegression().fit(df2_X_train, df2_y_train)

---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
<ipython-input-35-1b0e00e813dc> in <module>
----> 1 logistic_regression_model1 = LogisticRegression().fit(df2_X_train, df2_y_train)

~\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py in fit(self, X, y, sample_weight)
   1340             _dtype = [np.float64, np.float32]
   1341
-> 1342         X, y = self._validate_data(X, y, accept_sparse='csr', dtype=_dtype,
   1343                                    order="C",
   1344                                    accept_large_sparse=solver != 'liblinear')
```

One-Hot Encoding

currency	AUD	CAD	CHF	DKK	EUR	GBP	HKD	JPY	MXN	NOK	NZD	SEK	SGD	USD
GBP	0	0	0	0	0	1	0	0	0	0	0	0	0	0
USD	0	0	0	0	0	0	0	0	0	0	0	0	0	1
MXN	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Ordinal Encoding

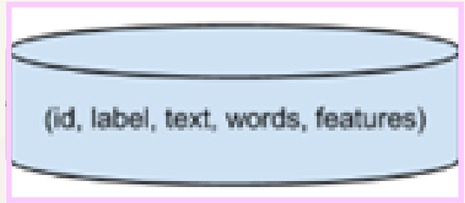
Dropped: High Multicollinearity with 'backers'

Dropped: Info Leakage

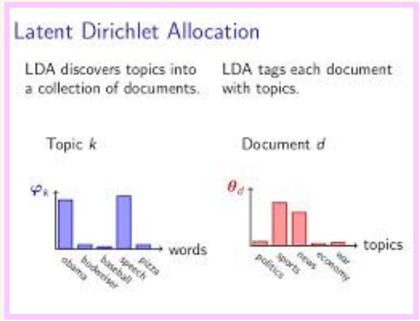
category	main_category	currency	backers	country	usd_pledged_real	usd_goal_real	campaign_duration_days
2	2	2	16	2	500.00	4500.00	44
155	11	2	22	2	460.00	400.00	29
22	9	3	96	3	25895.01	24656.05	29

NLP-DERIVED FEATURE TYPES

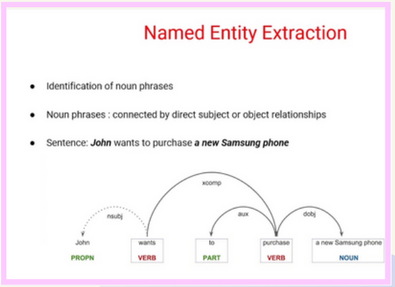
TEXT PROCESSING



TOPIC MODELING



NAMED ENTITIES



TEXT PREPROCESSING

Cleaning text for Topic Modeling

& calculating measurements on text

```
# Text Pre-Processing: Removing stopwords
# Remove stopwords from each Kickstarter's processed name
# 'remove_stopwords' is a custom function that takes in a string,
#      tokenizes it, and removes any stopwords in the string that
#      are in nltk's list of English stopwords.
df5['name_processed'] = df5['name_processed'].map(remove_stopwords)
```

	ID	name	name_processed	num_characters	caps_fraction	punctuations_fraction
46016	52842045	"Sir Clumsalot" Children's Book	sir clumsalot child book	31	0.129032	0.096774
61072	1190688864	Warm Quilts to Share	warm quilt share	20	0.150000	0.000000
51160	1270458613	Epirus: Tennis bags as stylish as the sport it...	epirus tennis bag stylish sport	50	0.040000	0.020000

PREPROCESSING STEPS ON KICKSTARTER TITLE:

- Convert title to all lowercase characters
- Remove URLs from title
- Remove mentions from title
- Remove numbers from title
- Remove punctuations from title
- Remove stopwords from title
- Lemmatize words in title

KICKSTARTER TITLE MEASUREMENTS:

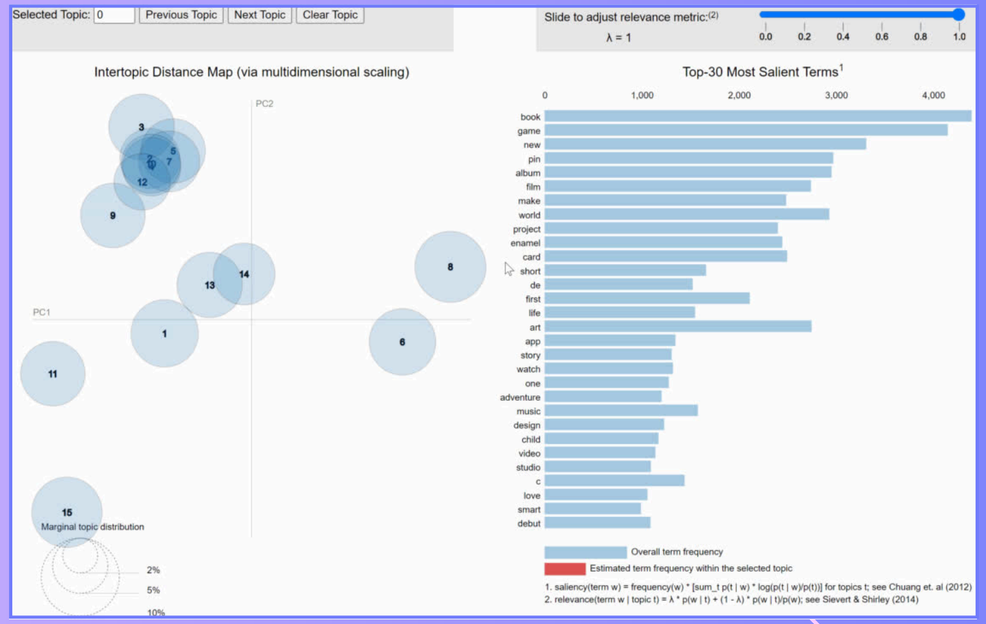
- Number of characters in title
- Number of capital letters / Total characters
- Number of punctuations / Total characters

TOPIC MODELING

Latent Dirichlet Allocation (LDA)

LDA as an unsupervised ML approach can group together related subtopics along with their corresponding word vector values. We can utilize this type of feature as a standalone variable that can enhance model performance. In addition, instead of solely designating one main topic, the clustered topics are assigned to specific documents which can later be used in root cause analysis to identify common patterns in topics associated with failed versus successful campaigns.

	ID	name_processed	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10	Topic_11	Topic_12	Topic_13	Topic_14	Topic_15
53368	1287371551	connection film	0.355498	0.022224	0.355587	0.022224	0.022224	0.022224	0.022224	0.022224	0.022224	0.022224	0.022224	0.022224	0.022224	0.022224	0.022224
67310	1302505626	pubbing	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667
69081	1936609685	greeting solar system	0.022222	0.022222	0.022222	0.022222	0.022222	0.022222	0.355550	0.022222	0.022222	0.022222	0.022222	0.355561	0.022222	0.022222	0.022222




```
Topic #15
Words in Topic:
        Word #1: 0.127*"game"
        Word #2: 0.051*"art"
        Word #3: 0.034*"c"
        Word #4: 0.030*"card"
        Word #5: 0.024*"board"
        Word #6: 0.023*"edition"
        Word #7: 0.021*"home"
        Word #8: 0.018*"rpg"
        Word #9: 0.017*"set"
        Word #10: 0.017*"high"
```

NAMED ENTITY RECOGNITION

Capturing & Labeling unstructured text into pre-defined categories

Examples of Labeled Entities in Kickstarter Titles

```
# NER Example 1
NER_ex1 = NER_output_list[15]
displacy.render(NER_ex1, style="ent", jupyter=True)
```

Destrus PERSON (FPS RPG ORG)

```
# NER Example 2
NER_ex2 = NER_output_list[45]
displacy.render(NER_ex2, style="ent", jupyter=True)
```

" Dead Tongues WORK_OF_ART " - Help Us Make a Lovecraftian Feature Film!

```
# NER Example 3
NER_ex3 = NER_output_list[1000]
displacy.render(NER_ex3, style="ent", jupyter=True)
```

DrewbyDoo Cosplay & Leatherworking ORG

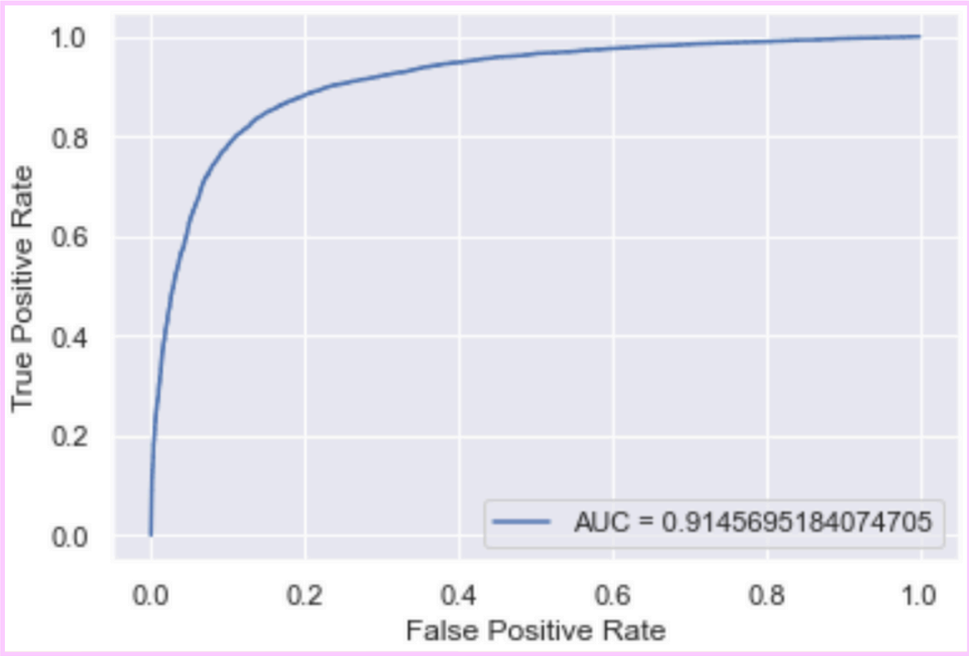
```
# NER Example 4
NER_ex4 = NER_output_list[5000]
displacy.render(NER_ex4, style="ent", jupyter=True)
```

Spartixed - FPGA Board ORG to learn Verilog / VHDL ORG

NER is an NLP practice often based off of supervised models trained on publicly available corpus datasets. We can use these labels to detect named entities, such as product and organization names, from the text and create variables that can prove relevant to predicting our target variable (e.g. WORK_OF_ART coinciding with the main_category "Music", which is overrepresented for successful campaigns).

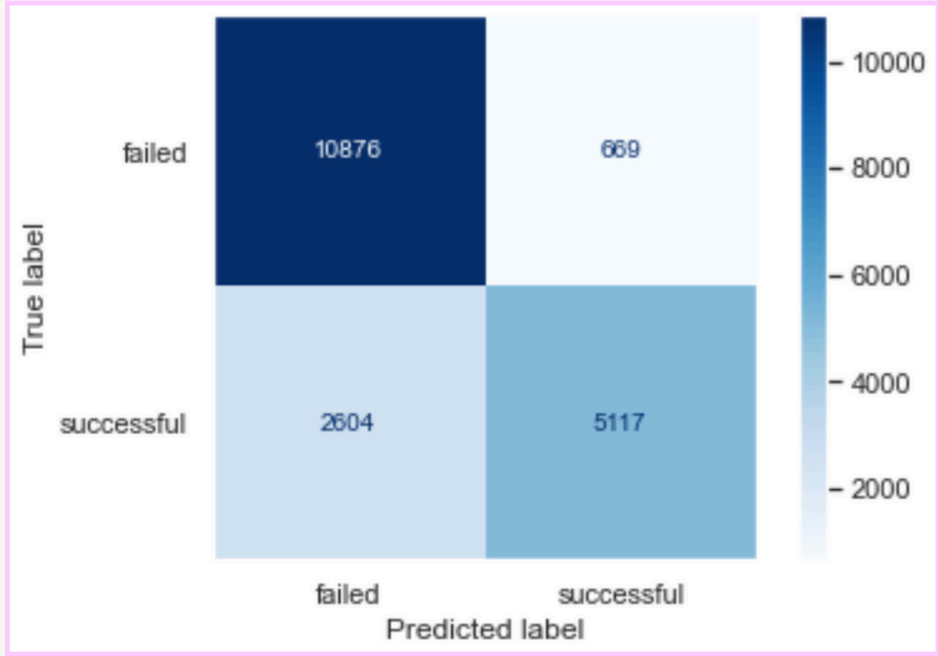
CARDINAL	DATE	EVENT	FAC	GPE	LANGUAGE	LAW	LOC	MONEY	NORP	ORDINAL	ORG	PERCENT	PERSON	PRODUCT	QUANTITY	TIME	WORK_OF_ART
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

FINAL LOGISTIC REGRESSION MODELING PERFORMANCE METRICS



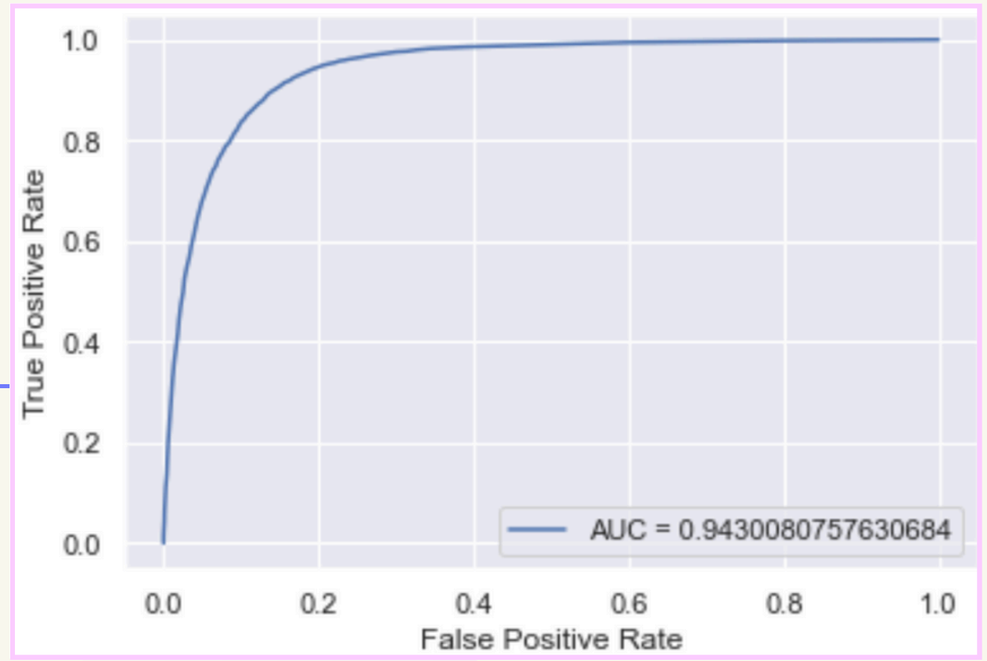
AUC

	precision	recall	f1-score	support
0	0.94	0.81	0.87	13480
1	0.66	0.88	0.76	5786
accuracy			0.83	19266
macro avg	0.80	0.85	0.81	19266
weighted avg	0.86	0.83	0.84	19266



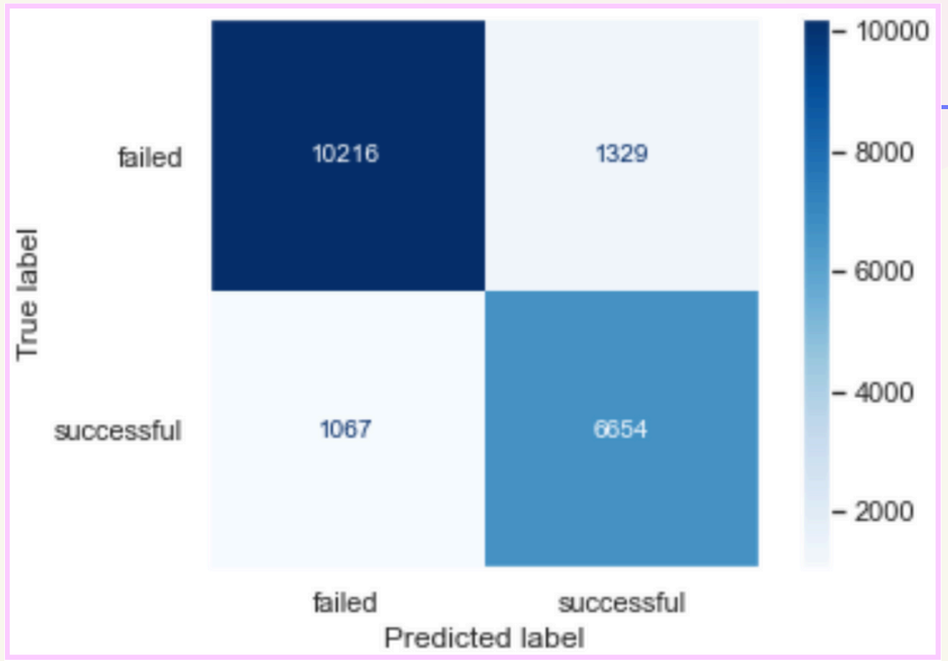
CONFUSION MATRIX

FINAL RANDOM FOREST MODELING PERFORMANCE METRICS



AUC

	precision	recall	f1-score	support
0	0.88	0.91	0.90	11283
1	0.86	0.83	0.85	7983
accuracy			0.88	19266
macro avg	0.87	0.87	0.87	19266
weighted avg	0.88	0.88	0.88	19266



CONFUSION MATRIX

THANK YOU