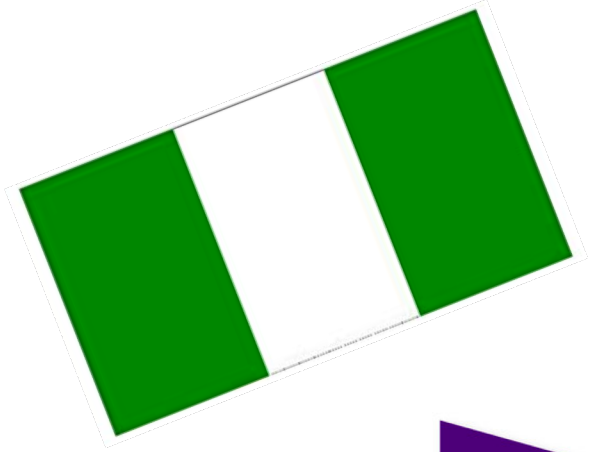
Machine Learning with SBA

Northwestern University Data Science Boot Camp | Sec. 2 |

July 17, 2018

Team

Francisco Galliano
Sunday Joseph



Special Thank you to

Our Families!

Our Faith & Priests

Our NU Cohorts

Final Project:

Wolf Bruckner, Sam Wood,

Abraham Eapen, Kevin Markman (Data Camp)

and Google – stack Overflow

Team Project

Overview

- Predict the amount of the Small Business Administration's (SBA) Guarantee based on State and Industries features
- Allow SBA to establish dynamic underwriting guidance based on the performance of the loan portfolio
- Provide a dashboard that can be used to visualize loan portfolio for the users discretion.

Data: SBA 7A Loans ➪ Data.gov

- Originally the dataset was over 200MB in size.
- Dataset contained over 1MM loan records.
- Loans from all 50 states plus territories from the period of 2010- 1Q2018.
- Categorical, Dates, Numerical, & Geographical
- 5 days - Analyzing, Clustering, Cleansing & etc.

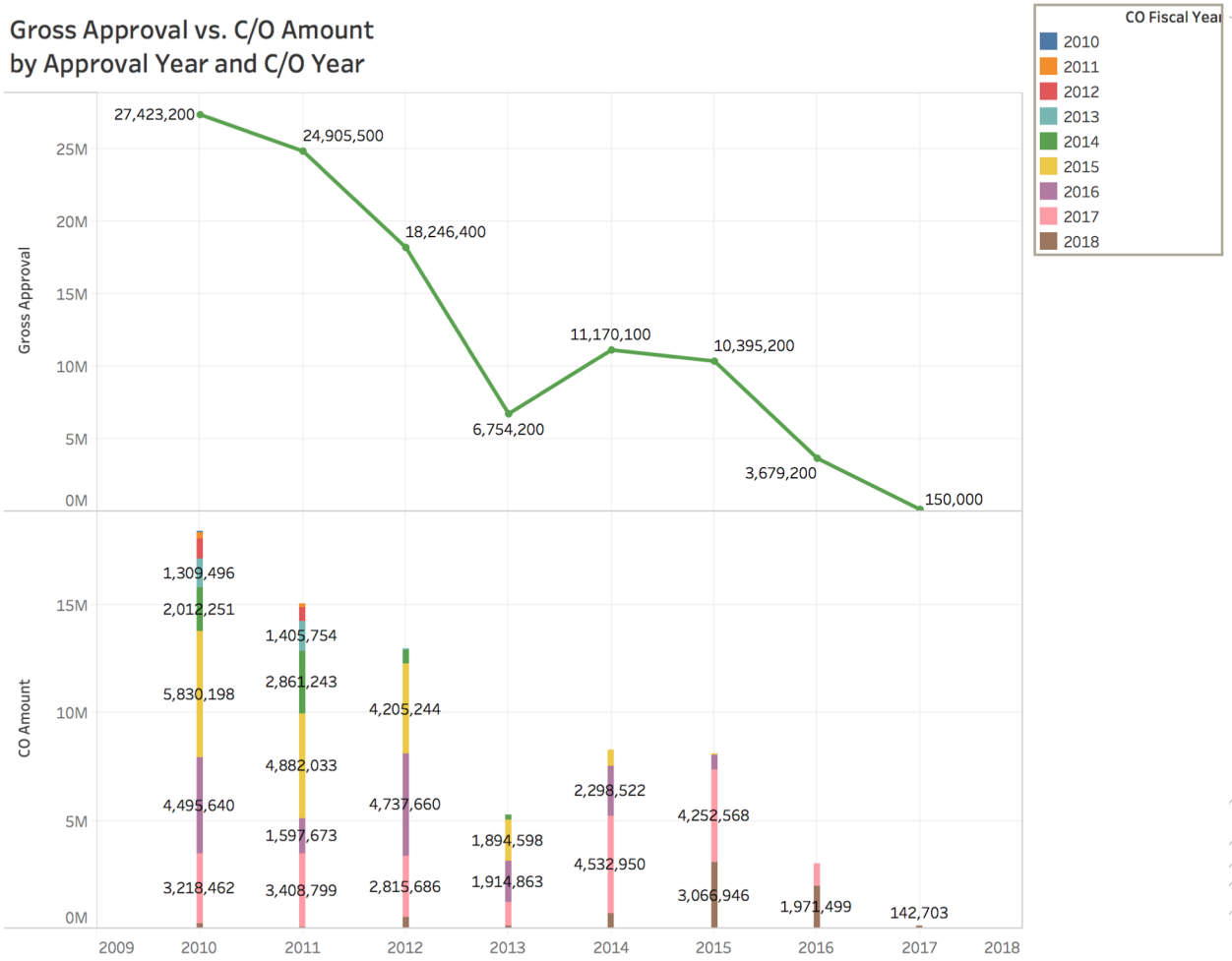
Challenges

- Noisy Data!
- Categorical Features ≠ Regression Analysis
- Model Selections
- Comprehending Correlations & Coefficients
- Analysis and structuring

BorrName	BorrStreet	BorrCity	BorrState	BorrZip	BorrID	LoanStatus	GrossApprov	SBAGuarante
TARCA, INC..	3754 NW 54	MIAMI	FL	33142	308	CHGOFF	1900000	1425000
Around the C	1242 SW Pin	Cape coral	FL	33991	8127	PIF	521800	469620
Integrated M	11701 South	Largo	FL	33773	4025	PIF	40800	34680
Loving Enter	7885 Boca Ci	SAINT PETER	FL	33706	7228	PIF	332000	249000
Lawrence J. I	249 Cataloni	Coral gables	FL	33134	1508	EXEMPT	225000	112500
KACEY HOME	2545 NORTH	TALLAHASSE	FL	32303	7816	PIF	800000	600000
William F. B	1250 S. Tiam	Sarasota	FL	34239	12966	PIF	75000	37500
Sergio A. Bal	13145 Old Cu	Miami	FL	33156	13037	PIF	50000	25000
ATLANTIC IN	2905 Danese	JACKSONVILI	FL	32206	5333	PIF	400000	360000
Wood Fired I	2235 Seacres	DELRAY BEA	FL	33444	10775	PIF	37600	18800
All Custom C	7656 NW 25	POMPANO B	FL	33063	10776	PIF	90000	45000
Aegis Fire ar	156 Industria	Orange Park	FL	32073	13132	PIF	200000	180000
Annette M. V	8910 Miramа	Miramar	FL	33025	1344	PIF	25000	22500
Flash Back D	220 South Fe	Hallandale	FL	33009	1483	EXEMPT	350000	175000
A F S Carrier	14250 SW 35	Miami	FL	33175	1506	EXEMPT	200000	100000
ATLANTIC HE	3960 WILLOV	PORT ORAN	FL	32127	6776	PIF	50000	25000
Educare Acad	10220 West	Bonita Spring	FL	34135	7649	EXEMPT	725000	652500
The Best Foc	1170 NE 191	MIAMI	FL	33179	10771	PIF	25000	22500
Viking Diving	2899 SE Grar	PORT SAINT	FL	34952	10789	PIF	50000	42500
Boca Extrem	22954 Ironw	BOCA RATON	FL	33433	10790	PIF	25000	12500
HALINA'S CU	5346 MAIN S	NEW PORT F	FL	34652	13568	EXEMPT	15000	13500
ADAPTIVE LI	2473 EGRETS	JACKSONVILI	FL	32224	13593	EXEMPT	10000	9000
JOHN A ALLN	4661 KERLE S	JACKSONVILI	FL	32205	14395	PIF	10000	5000
LEGACY WEA	7890 PETERS	PLANTATION	FL	33324	15302	EXEMPT	150000	127500
Aldo M. Leiva	9155 Dadela	MIAMI	FL	33156	16472	PIF	40000	20000

Dashboard - Tableau

One of the initial observations at we noticed was that loan charge-offs attend to peak at year 4-5 after approval. The dashboard on the right is for Florida. $10.3MM was c/o in year 2015 & 16 for loans funded in 2010. For the 2011 pool, $7.7MM was c/o in 2014 & 15.



Machine Learning

- At the beginning it was determined that we had enough information(features) for a RA model to test and determine what good loans would go bad. Well that didn't go so well!
- When the model could not provide a status, we attempted to predict a charge-off amount for good loans. That too burned up at take off!
- Noticed that the dataset was very noisy.
- Eliminated all Canceled, unfunded loans & loans with a negative C/O Amount.

Frustration Ensues!!

Correlation Analysis

Machine Learning Cont.

- Performed 4 regression analyzes – Linear, Lasso, Ridge and ElasticNet
- Due to file size we sliced the data by states: IL, FL, NY & TX.
- Then create clusters to learn how the models would perform across a sample of industries.
- Our models performed by when we set the Random State to 60.
- Models were very sensitive to small samples.

Let's Bring It Home!

The Team's

Models & Results

- STATE of FLORIDA REGRESSIONAL

- Florida_Model_4_Industries

Conclusion

The results:

Our predicted values are not perfect and there is some level of error, but we have a solid base from which to build from for version 2.0

The challenges:

Team size and data features

Any interesting/insightful:

Can build a predictive model in 7 days. ”For Loops” are so awesome. GitPages are great for developing a static webpage.

Questions?

Thank you and lets get the heck out of here!!

DELIVERABLE ITEMS FOR PROJECT

Project Requirements @ 6/30/18:

Problem Worth Solving ☑
Analyze ☑
Visualize ☑

Utilize:
- Sci-Kit Learn ☑

Use at Least Two Tools: ☑
- Python Pandas
- Python Matplotlib ☑
- Tableau ☑

Wolf's Slack Message on 7/14/18

For your presentations you should discuss:

- Your data ☑
- What you set out to show with your data ☑
- The results you got, if they match your goals from above (or if they dont, why) ☑
- What challenges you faced while working on your project ☑
- Any interesting/insightful things you found while working (helpful libraries, coding tips/tricks) ☑