Small Business Loans: Predicting Default

Improving The Small Business Administration Loan Commitment Program

Dillan Gump

Context: Small Business Administration Loan Program

- SBA sponsors a program that commits to guarantee a portion of a loan
- If a loan defaults, the SBA is responsible for the portion committed
- If only there was a way to know if a loan would default ahead of time...

Purpose - For the SBA

- Help the SBA decide which loans to approve
- Determine the factors that contribute to a loan defaulting
- Improve upon traditional selection techniques with machine learning

What will a good model look like?

- Approves as many loans as possible while minimizing risk
  - Manage backing loans that are likely to default
- Prioritizes incorrectly predicting default over incorrectly predicting pay-off
  - Conservative model
- Sacrifices and Trade-Offs
  - May capture too many loans that won't default
  - Big Con. Can be tuned.

What We Know

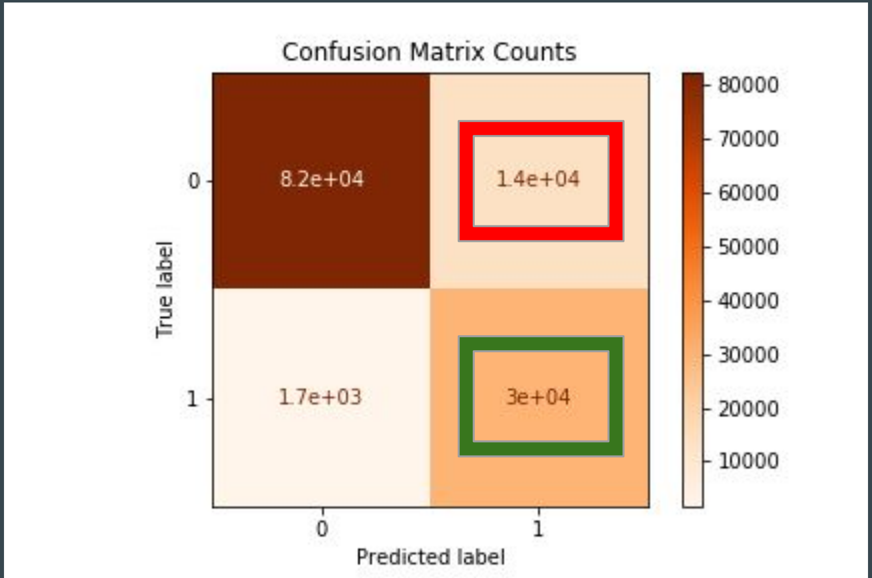
- Original Data
  - ~800,000 loans
  - 1965 - 2014
- Selected Data
  - ~500,000 loans
  - 2000 - 2014
  - Completed Loans Only
- Features
  - Loan Information
  - Bank Information
  - Business Information
  - SBA information

New Information

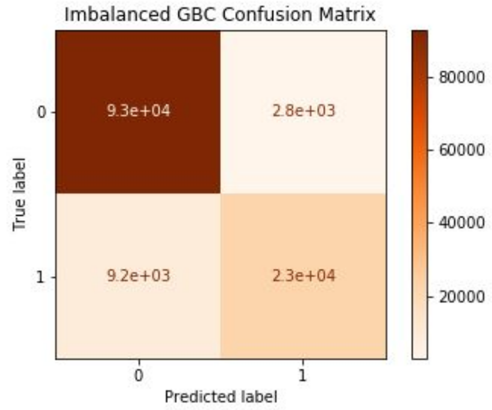
- Using the existing data, a few new features were created
- New Features:
  - Yes/No: Is the business a franchise?
  - Yes/No: Is the lending bank out of state?
  - Yes/No: Was the loan approved by the SBA before disbursement?
  - Levels: Number of jobs created and number of jobs retained
  - Levels: Bank size
  - Percentage: Portion of loan covered by SBA
- Good start towards model improvement

Model Summary

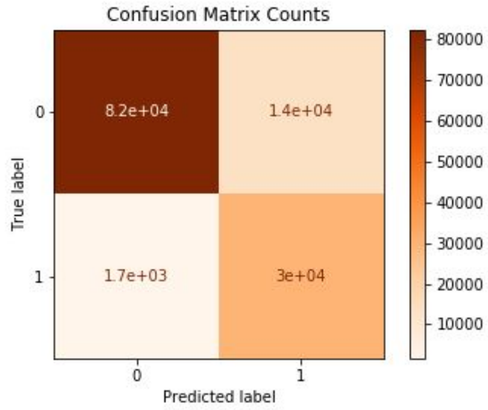
- Accuracy: 0.88
- Precision: 0.68
  - 32% of predicted defaults were wrong
- Recall: 0.95
  - 95% of defaulted loans were correctly identified



Imbalanced Classifier Comparison - Tuning Demo



	precision	recall	f1-score	support
0	0.91	0.97	0.94	95775
1	0.89	0.71	0.79	31765
accuracy			0.91	127540
macro avg	0.90	0.84	0.87	127540
weighted avg	0.91	0.91	0.90	127540



	precision	recall	f1-score	support
0	0.98	0.85	0.91	96230
1	0.68	0.95	0.79	31713
accuracy			0.88	127943
macro avg	0.83	0.90	0.85	127943
weighted avg	0.91	0.88	0.88	127943

```python
preprocessing = ColumnTransformer(
    [
        ("leaveOneOut", LeaveOneOutEncoder(), cat_cols),
        ("scale", StandardScaler(), num_cols),  # never hurts
        # ("knnImptute", KNNImputer(n_neighbors=2), impute_cols),
        # ("simpleImptute", SimpleImputer(), impute_cols),
    ],
    remainder="passthrough",
)
```

```python
n_trees = 100
learning_rate = 2 / n_trees

pipeline = Pipeline(
    [
        ("preprocessing", preprocessing),
        ("xgbClass", XGBClassifier(n_estimators=n_trees, learning_rate=learning_rate)),
    ]
)
```

```python
grid = {
    "xgbClass__subsample": [0.00125,0.0025, 0.01],
    #     "gbr__max_features": [0.5, 0.75, 1.0], # alternative
    "xgbClass__colsample_bytree": [0.6, 0.8, 1.0],
    "xgbClass__max_depth": [4, 6,7,8],
}


```

```python
pipeline_cv_reclean.best_params_
```

```
{'xgbClass__colsample_bytree': 1.0,
 'xgbClass__max_depth': 6,
 'xgbClass__subsample': 0.01}
```

```python
resample_grid = {
    "xgbClass__subsample": [0.1, 0.5],
    "xgbClass__colsample_bytree": [0.6, 0.8, 1.0],
    "xgbClass__max_depth": [4, 6,7,8],
}

```

```python
pipeline_cv_resample_reclean.best_params_
```

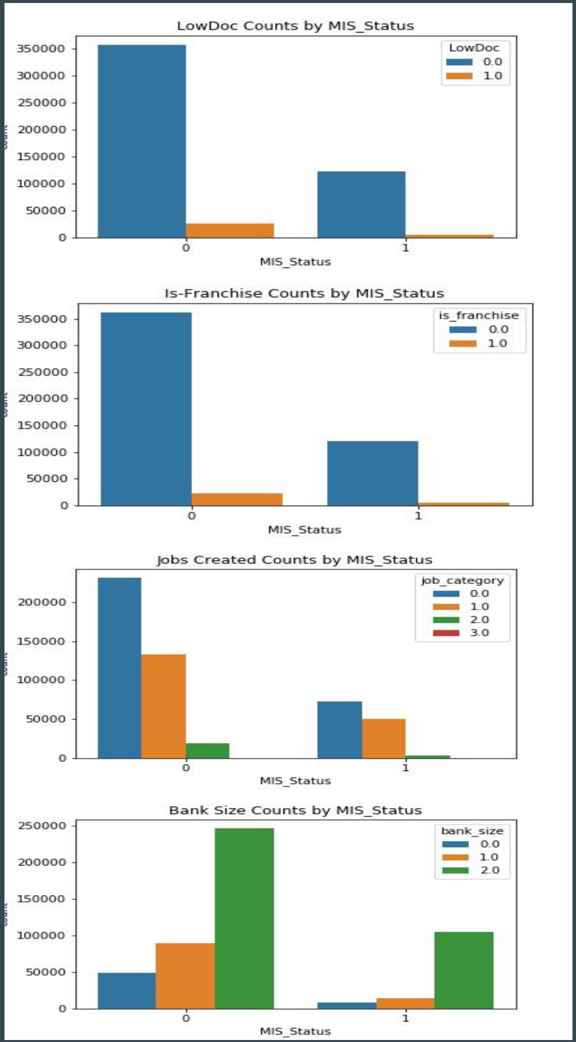
```
{'xgbClass__colsample_bytree': 1.0,
 'xgbClass__max_depth': 7,
 'xgbClass__subsample': 0.1}
```

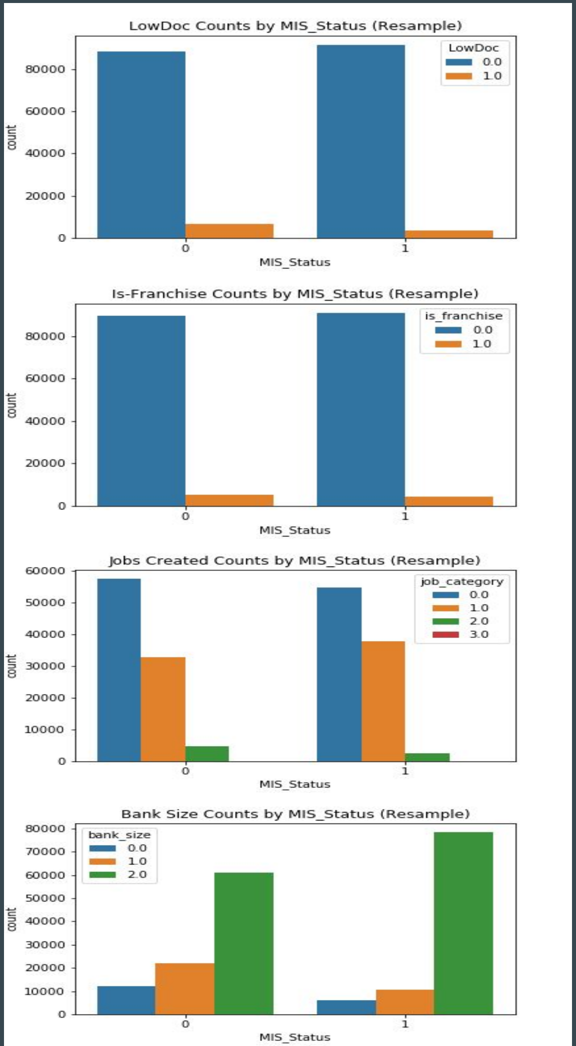
Resampling Technique

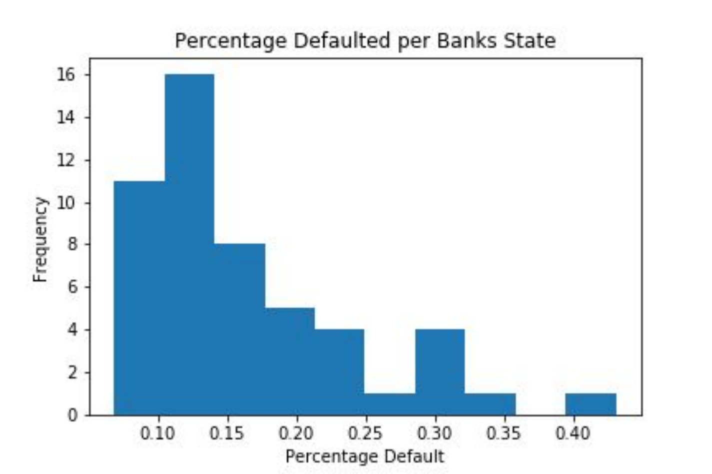
```python
X_train_0 = X_train[y_train == 0]
X_train_1 = X_train[y_train == 1]

n_0 = X_train_0.shape[0]
n_1 = X_train_1.shape[0]

# Sample majority class to have less observations
X_train_0_sample = X_train_0.sample(n_1, replace=False, random_state=42)

# # Sample minority class to have less observations
# X_train_1_sample = X_train_1.sample(n, replace=True, random_state=42)

X_train_resample = pd.concat((X_train_1, X_train_0_sample))
X_train_resample = X_train_resample.reset_index(drop=True)

y_train_resample = np.array([1] * n_1 + [0] * n_1)
y_train_resample.mean()
```

Important Features

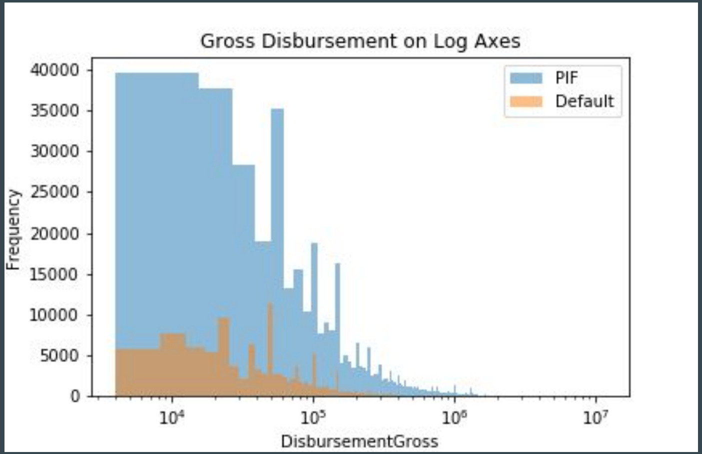
- Low Documentation (Y/N): Whether a loan under $150k can complete an alternative 1-page application
    - Strongest predictor. Makes sense as it is a kind of prescreening
- Job Category(Levels): How many jobs were created
    - Loans in the 10-100 jobs created range had the smallest percentage default
- Bank State (Categorical): State of bank issuing loan
- Disbursement Gross (Continuous): Defaulted loans tended to have a smaller sum
    - Mean and median both low
- Is Franchise (Y/N): Is the business a franchise?
- Bank Size (Levels): Based on number of loans given


LowDoc Counts by MIS_Status
LowDoc
0.0
1.0
count
MIS_Status
Is-Franchise Counts by MIS_Status
is_franchise
0.0
1.0
count
MIS_Status
Jobs Created Counts by MIS_Status
job_category
0.0
1.0
2.0
3.0
count
MIS_Status
Bank Size Counts by MIS_Status
bank_size
0.0
1.0
2.0
count
MIS_Status


LowDoc Counts by MIS_Status (Resample)
LowDoc
0.0
1.0
count
MIS_Status
Is-Franchise Counts by MIS_Status (Resample)
is_franchise
0.0
1.0
count
MIS_Status
Jobs Created Counts by MIS_Status (Resample)
job_category
0.0
1.0
2.0
3.0
count
MIS_Status
Bank Size Counts by MIS_Status (Resample)
bank_size
0.0
1.0
2.0
count
MIS_Status



BankState	
VA	0.430738
NC	0.339468
SD	0.316343
CA	0.307649
OR	0.293199
IL	0.291240
DE	0.285044
FL	0.244540
NY	0.237958
RI	0.232725
UT	0.219884
OH	0.204191
SC	0.202687
AL	0.199076
CT	0.190476
TX	0.180539
MO	0.175309
WI	0.168798
WV	0.167010
DC	0.162338
IA	0.162050
AR	0.161411
KY	0.160494
GA	0.159182
ID	0.139535
NE	0.138377
OK	0.137051
AK	0.134615
TN	0.130936
MS	0.128300
NJ	0.126822
HI	0.126208
NV	0.125000
MI	0.123644
KS	0.123408
MD	0.120241
WA	0.117607
MN	0.114859
IN	0.112768
LA	0.107488
NM	0.102632
PA	0.096965
ME	0.095610
MA	0.094795
AZ	0.089459
ND	0.083294
NH	0.082742
CO	0.076905
MT	0.069466
VT	0.069134
WY	0.068323



Final Interpretation

- Existing signs of confidence are good indicators of loan success
  - Approved for Low Doc, growing employees
- Not really any harm in losing precision
  - Those loans ended up paid anyways
- The data shows a lot of potential
  - True effect of state
  - Refine engineered features
- Future Improvements
  - Predict how much a loan defaulted
  - Further Subsetting
  - Tune Results

			DisbursementGross	
			mean	median
LowDoc	MIS_Status	bank_size		
No	Defaulted	Large	94823.800000	50000.000000
		Medium	244785.450000	135500.000000
		Small	235739.170000	145000.000000
	PIF	Large	165073.970000	57130.500000
		Medium	331792.630000	182000.000000
		Small	317841.520000	167161.000000
Yes	Defaulted	Large	91940.020000	94869.500000
		Medium	93476.110000	92000.000000
		Small	89542.970000	90000.000000
	PIF	Large	87747.750000	85000.000000
		Medium	86167.110000	81600.000000
		Small	82967.640000	75000.000000

LowDoc	MIS_Status	bank_size	DisbursementGross mean	DisbursementGross median
No	Defaulted	Large	94823.800000	50000.000000
		Medium	244785.450000	135500.000000
		Small	235739.170000	145000.000000
	PIF	Large	165073.970000	57130.500000
		Medium	331792.630000	182000.000000
		Small	317841.520000	167161.000000
Yes	Defaulted	Large	91940.020000	94869.500000
		Medium	93476.110000	92000.000000
		Small	89542.970000	90000.000000
	PIF	Large	87747.750000	85000.000000
		Medium	86167.110000	81600.000000
		Small	82967.640000	75000.000000

LowDoc	MIS_Status	job_category	DisbursementGross mean	DisbursementGross median
No	Defaulted	0.0	117750.370000	50000.000000
		1.0	103052.660000	50000.000000
		2.0	294019.550000	175500.000000
		3.0	356571.190000	200992.000000
	PIF	0.0	193336.890000	70732.000000
		1.0	209639.060000	100000.000000
		2.0	547624.910000	406000.000000
		3.0	843321.020000	542250.000000
Yes	Defaulted	0.0	91506.060000	90000.000000
	PIF	0.0	85278.410000	80000.000000

Feature	Importance
LowDoc	0.280424
job_category	0.133911
BankState	0.078828
DisbursementGross	0.055349
is_franchise	0.051034
bank_size	0.047374
State	0.042923
RevLineCr	0.034007
bank_out_of_state	0.032604
UrbanRural_cleaned	0.031403
Disbr_year	0.030572
twoDigNAICS	0.028628
NewExist	0.027750
retained_category	0.025924
NoEmp	0.025847
Term_years	0.025280
Disbr_Month_cos	0.024222
Disbr_Month_sin	0.023920

LowDoc	0.0	1.0
MIS_Status		
0	0.746469	0.847128
1	0.253531	0.152872

	DisbursementGross	
	mean	median
LowDoc		
0.0	192753.906139	75000.0
1.0	86988.485540	82300.0

	DisbursementGross	
	mean	median
MIS_Status		
0	209783.519929	87300.0
1	115712.153896	51437.0

		DisbursementGross	
		mean	median
LowDoc	MIS_Status		
0.0	0	218678.793757	88599.0
	1	116630.470726	50000.0
1.0	0	86153.722524	80000.0
	1	91616.073843	90000.0

	DisbursementGross	
	mean	median
bank_size		
0.0	260532.150587	132000.0
1.0	294007.328827	150000.0
2.0	142883.929899	51243.5

job_category	0.0	1.0	2.0	3.0
MIS_Status				
0	0.760818	0.725964	0.843943	0.76584
1	0.239182	0.274036	0.156057	0.23416

is_franchise	0.0	1.0
MIS_Status		
0	0.749927	0.796707
1	0.250073	0.203293

	DisbursementGross	
	mean	median
job_category		
0.0	166810.504676	66808.0
1.0	178753.683634	79100.0
2.0	506060.262202	350000.0
3.0	601604.148760	393000.0

Implementation

- Long Term Validation Study
- Deploy and compare several new strategies
  - Control- Traditional methods
  - Other models built with different information
  - Collaboration with business team
  - Introduce new data, engineer new features
- Integrate model as a software tool
  - Possible automation
  - Augment Existing Process
- Supplement with Diversity and Inclusion Measures
  - Model won’t challenge historical trends

Questions?

Thank You!

Appendix

- SBA Link: https://www.sba.gov/offices/headquarters/ofa/resources/11421
- Link to original Kaggle set: https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied
-