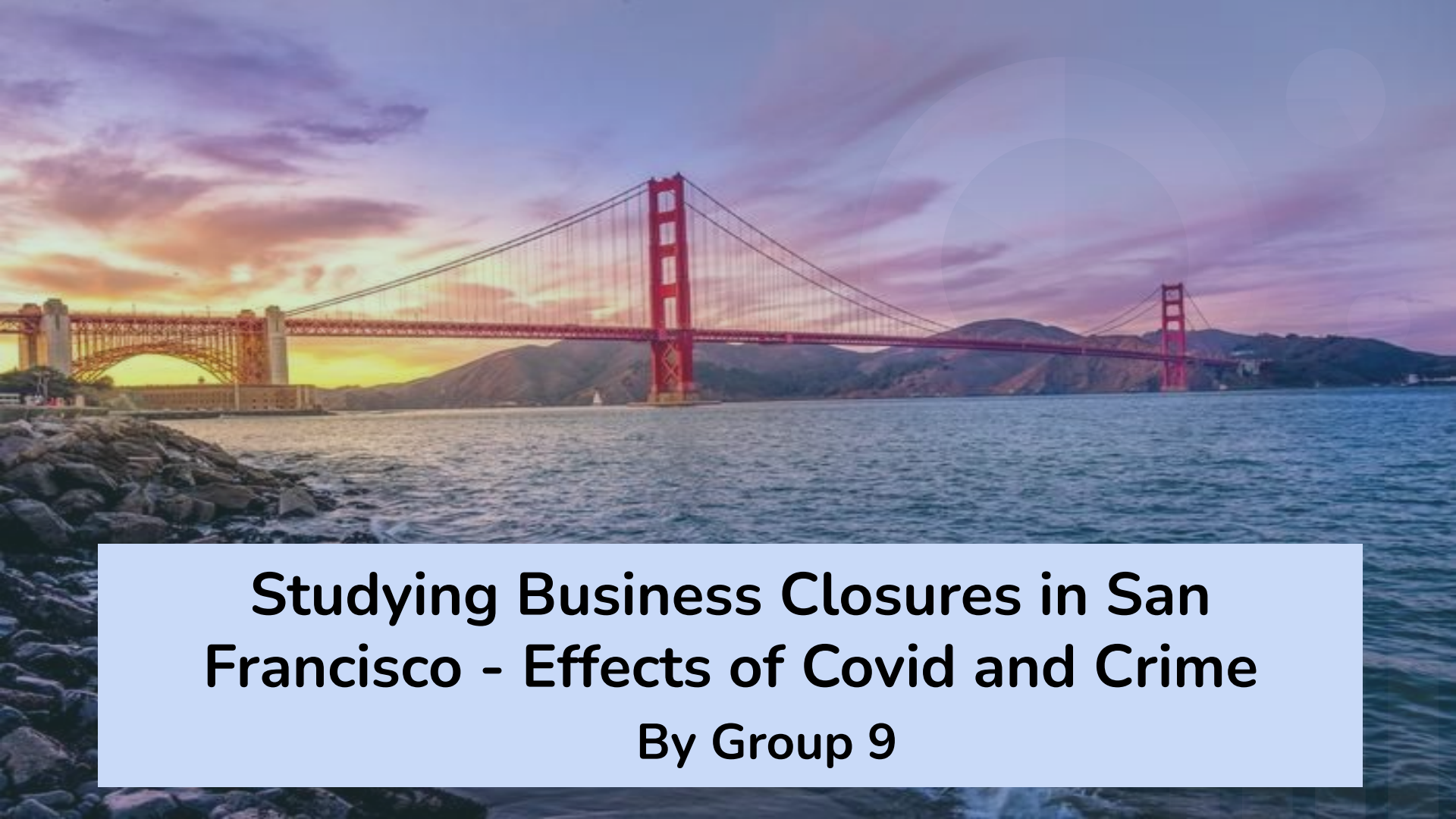


Studying Business Closures in San Francisco - Effects of Covid and Crime

By Group 9

Analytical Goals

- Combine different data sources to understand the influence of crime incidents, Geography, covid cases, vaccinations on business closures at a location in San Francisco.
- Build Machine Learning Models to predict whether a business at a location will closedown in the next six months based on different attributes described above.

Data and Infrastructure

Data and Sources

1. Police_Department_Incident_Reports: Date wise, Geography wise crime incidents reported
2. COVID-19_Cases_by_Geography_and_Date:
3. Registered businesses: Geography wise Business name, description, business start date and permanent closer date etc
4. Area Walk scores

Sources:

- https://data.sfgov.org/
- https://www.walkscore.com/CA/San_Francisco

Data Cleaning and Preprocessing

- Time period considered for the Analysis June 2018 to Dec 2021
- San Francisco is geographically divided into 41 neighborhoods. Used these neighbourhoods to join different data tables available.
- Created a flag for a business to indicate whether it got closed at the location during the following six months.
- Created monthly covid and crime counts from daily data.
- Preprocessed the data on databricks using Spark cluster.
- Post clean up loaded the data into MongoDB to store the data.

Task 2 - Cluster setting and Preprocessing Time

- **Cluster :** 3 Node i3.xlarge cluster with v.10.3 (includes Apache Spark 3.2.1, Scala 2.12)
- **Preprocessing Time:** 415.8 secs


mirror_mod = modifier_ob.
mirror object to mirror
mirror_mod.mirror_object
operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True
selection at the end -ad
_ob.select= 1
er_ob.select=1
ntext.scene.objects.acti
("Selected" + str(modifier
irror_ob.select = 0
bpy.context.selected_ob
ata.objects[one.name].se
int("please select exact
OPERATOR CLASSES

ML Models

Features generated for model development

S. No.	Columns	Description
1	Acs_population	Total Population
2	Flag_vaccin_gt75	Binary showing if CA vaccination rate is more than 75%
3	Business_age	Age of the business
4	Walk_Score	Score depicting ease of walking
5	Transit_Score	Score depicting ease and availability of public transit
6	Bike score	Score depicting ease of biking in an area
7	sum_%_pop_with_covid_6m	Percent population with active covid in last 6 months
8	Sum_covid_cases_6m	Total number of covid cases in last 6 months
9	Sum_crime_count_6m	Total number of crime cases in last 6 months
10	AnalysisNeighborhood	Neighborhood in San Francisco
11	NAICS_code_desc	Industry of the business
12	Police_district, current_super_dist, supervisor_dist	Location Identifier

Data Transformations

StringIndexer	Used StringIndexer to convert the columns with String data type to numeric.
OneHotEncoder	Used OneHotEncoder to encode the numeric versions of the String columns
Vector Assembler	Assembled the all the feature variables into a feature vector using Vector Assembler
Train/Test Split	Split the dataset into 80% train rows and 20% test rows

ML Algorithms Comparison

Performance comparison of the different models developed to predict business closure is as below:

Sr. No.	ML Algorithm	Best Model Parameters	F1 Score	Execution Time
1	Logistic Regression CV	regParam: 0.1	0.88	805.81 sec
2	Decision Tree	MaxDepth : 20	0.93	1366.64 sec
3	Random Forest	MaxDepth : 20	0.89	358 sec
4	Stacked Ensemble Model	MaxDepth : 20, GB iter : 20, learning rate : 0.005	0.81	2262.38 sec

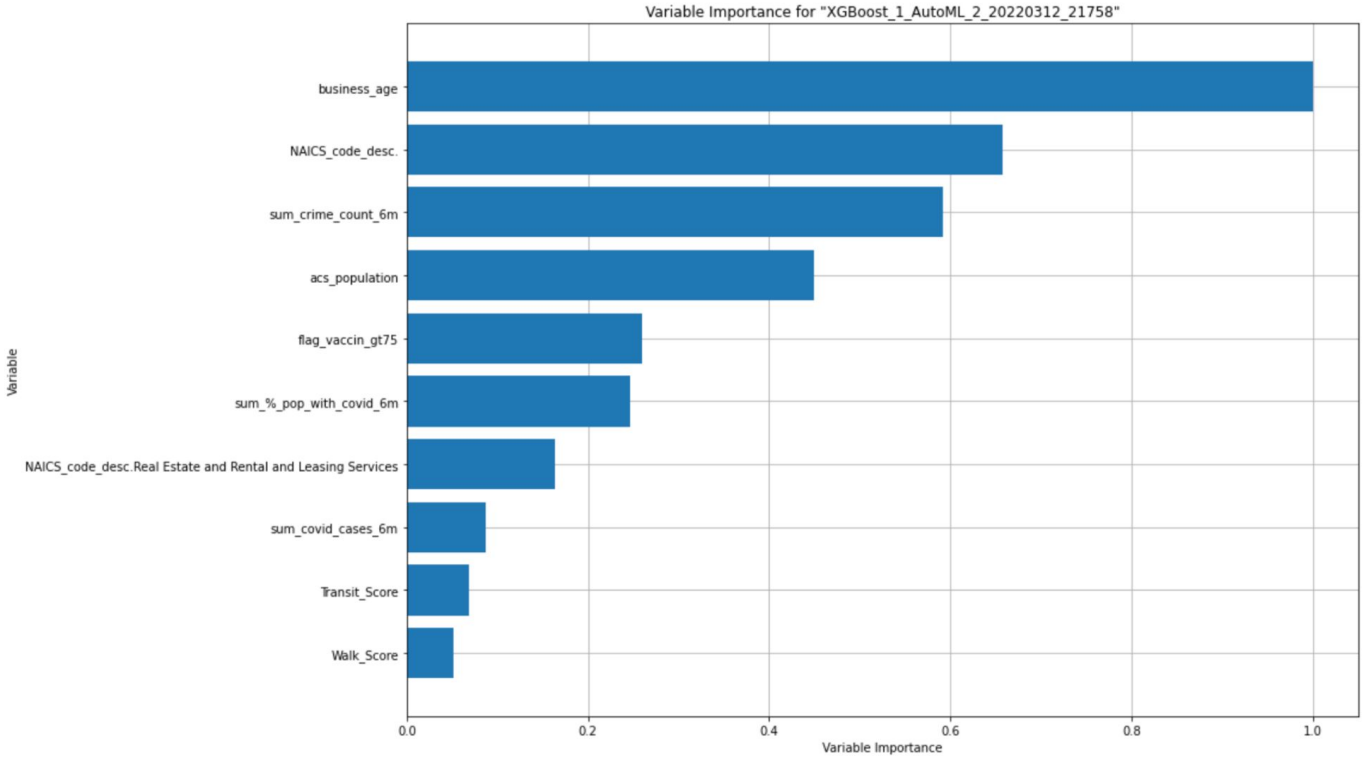
AutoML

Used AutoMl to find the best model. AutoMl helps to choose the best model that fits a dataset and also explains feature importances and other parameters along with dependency plots. The leaderboard of the top 3 models are shown below.

Out[225]:

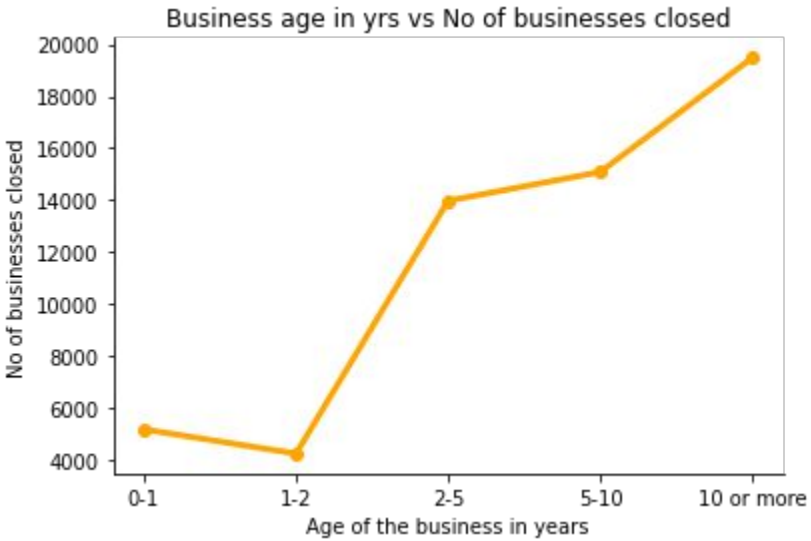
model_id	auc	logloss	aucpr	-
XGBoost_1_AutoML_2_20220312_21758	0.761981	0.21465	0.293392	
GBM_1_AutoML_2_20220312_21758	0.750803	0.23308	0.267395	
GLM_1_AutoML_2_20220312_21758	0.703215	0.238719	0.148205	

Feature importances from the best AutoML model



Top 3 Features - Deep dive

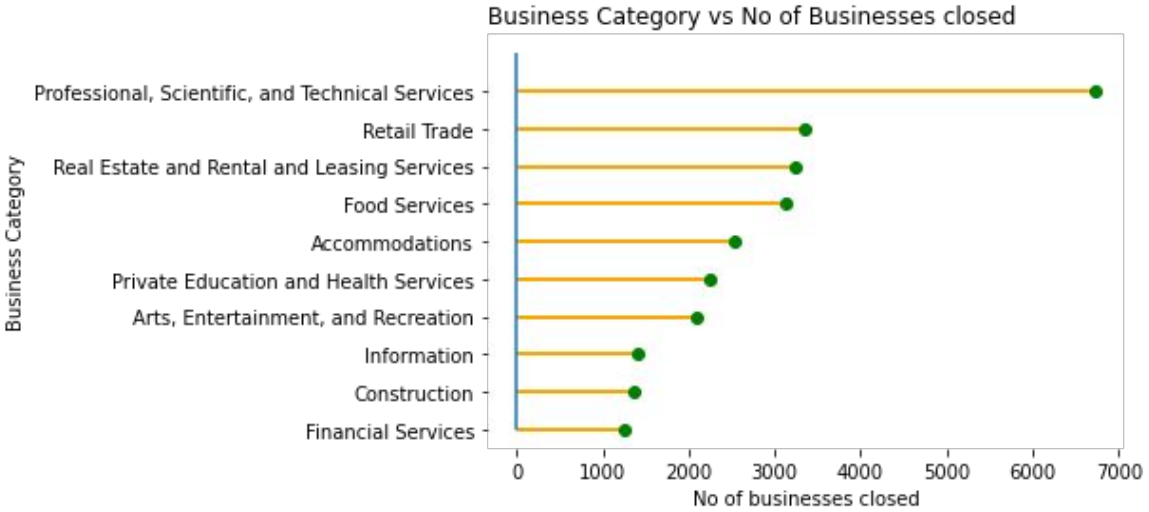
Age of the Business in years Vs Business Closure



Age of the Business is consistently identified as an Important feature influencing business closures.

Significantly More No of older (more than 5yrs) Businesses closed during the period as compared to younger businesses

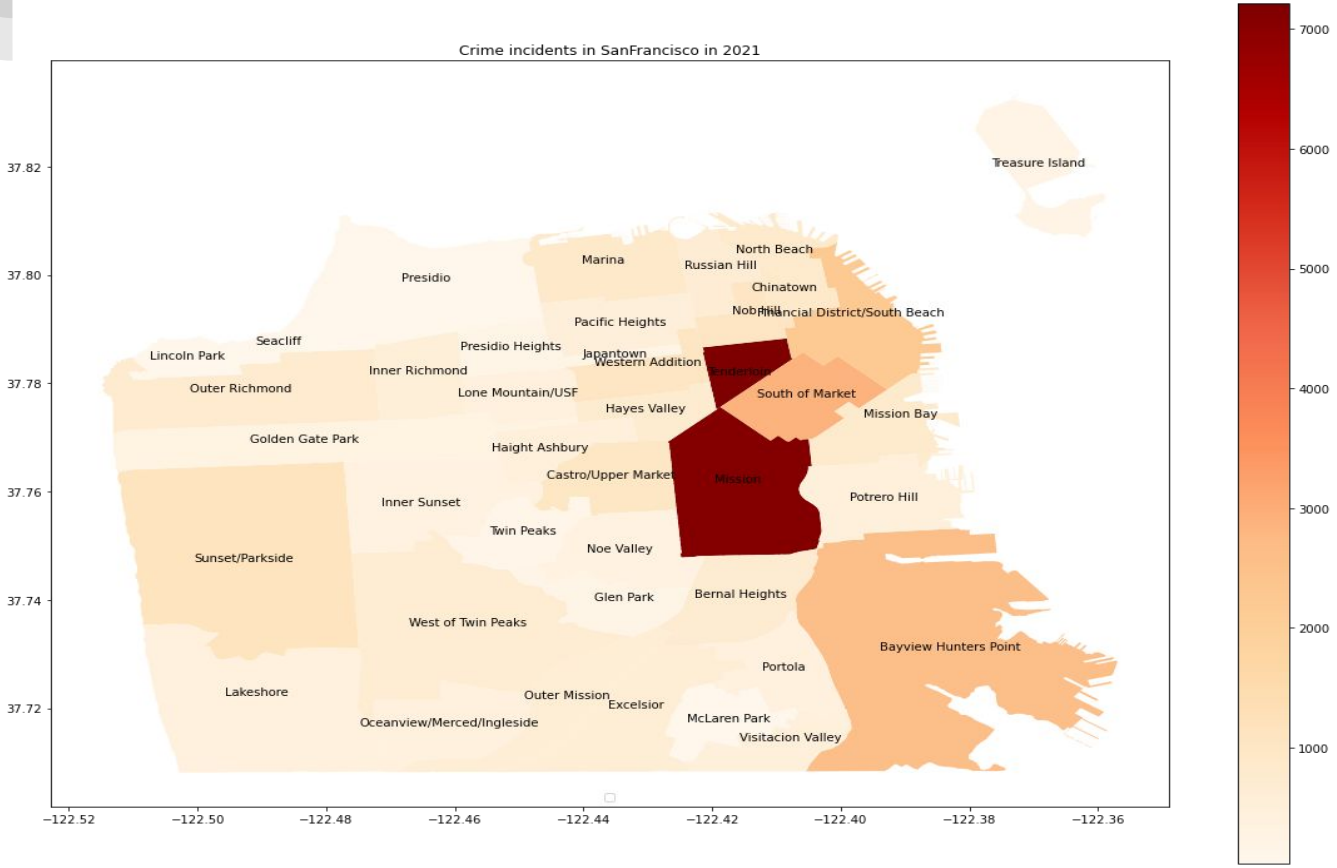
Type of Businesses closed



Over 20000 businesses closed have no NAICS-code.

Among those available, Top 10 nature of businesses closed is given here.

Area wise crime incidents in San Francisco in 2021



Task 3 - Cluster setting

- **Cluster** : i3.xlarge cluster with v.10.3 (includes Apache Spark 3.2.1, Scala 2.12)
  - 2 to 8 workers
  - 30.5 GB memory
  - 4 Cores
- **Packages Installed** : h2o_pysparkling_3.2

databricks™

Conclusions

- The main factors affecting the business activity in San Francisco are :
  - Nature of the business: Retail, Food services and accommodations businesses were impacted negatively during Covid.
  - Age of the Business: More Older(5+yrs) businesses have closed during the 2018-21 time period.
  - Crime rates in an area is a significant predictor of business closure.
- Active covid cases, vaccination rates did not have a significant impact on closing of a business.

Lessons Learned

- Huge datasets can be integrated and processed using clusters and distributed computing on Databricks.
- MongoDB can be used effectively for storing, loading and querying big unstructured datasets.
- SparkML can run machine learning algorithms using distributed computing.
- Automated machine learning is the process of automating the tasks of applying machine learning algorithms. It fits data using multiple models and ranks them in the order of a desired metric (AUC, R2 etc).


Thank
you

Appendix: Model Details

Logistic Regression Model

- Used a cross validation loop to train a logistic regression model.
- Hyper Parameters tuned:
  - regParam: [0.0001, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5]
- Metrics:
  - AUC: 0.7061
  - F1: 0.883
  - AreaUnderPR: 0.1326
- Model performs poorly whale predicting 0’s

```
Confusion Metrics =
DenseMatrix([[39327.,     0.],
             [ 2816.,     0.]])
```

AutoML - undersampled 0’s data

Used AutoMl for dataset with undersampled 0’s. StackedEnsemble models worked the best.

model_id	auc	logloss	aucpr
StackedEnsemble_AllModels_1_AutoML_1_20220312_50006	0.768569	0.478112	0.603774
StackedEnsemble_BestOfFamily_1_AutoML_1_20220312_50006	0.768242	0.478295	0.603778
StackedEnsemble_BestOfFamily_2_AutoML_1_20220312_50006	0.767133	0.479478	0.60285
GBM_1_AutoML_1_20220312_50006	0.766705	0.479484	0.601904
XGBoost_1_AutoML_1_20220312_50006	0.761699	0.485042	0.594535

Decision Tree

- Hyper parameter Tuning: Used Cross validation for Tree depth
- HyperParameters Tuned: [5,10,15,20,25,30]
- **Best Model:** max_depth=20
- **Metrics:**

  F1 Score: 0.92

- ```
  Confusion Metrics =
  DenseMatrix([[155783.,   1356.],
               [  8496.,   3162.]])
  ```
  ```

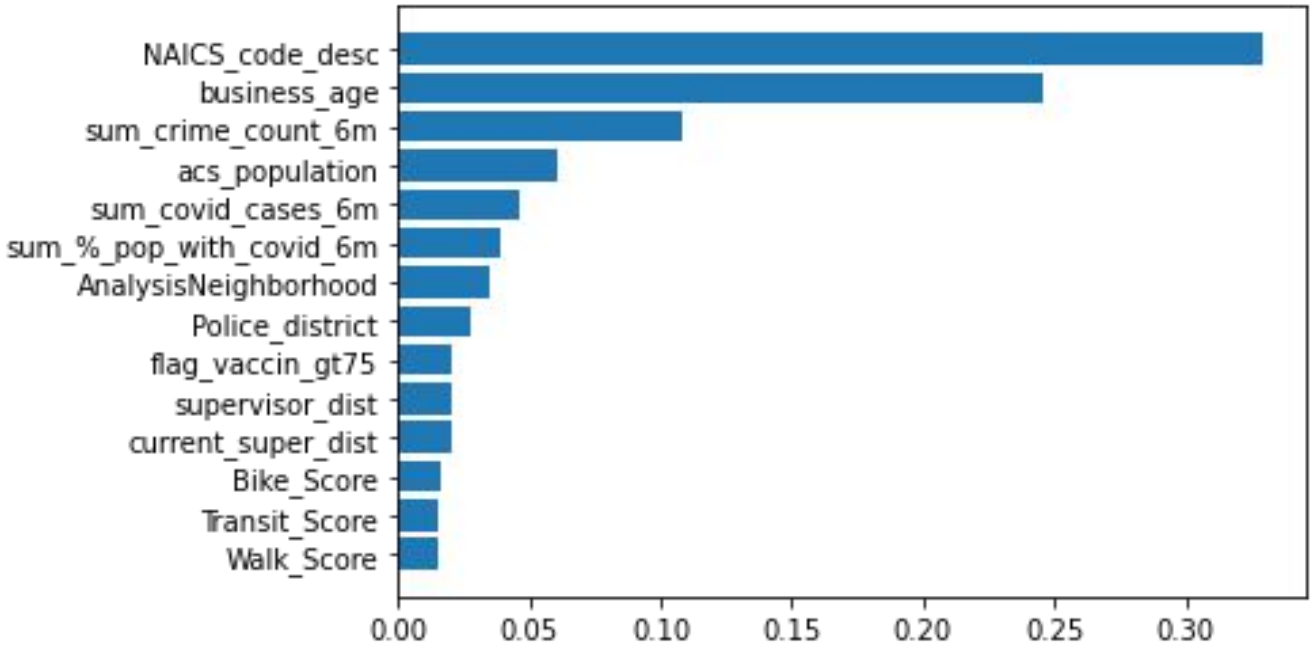
Random Forest

- Hyper parameter Tuning: Used Cross validation for Tree depth,
- Manually tuned for giving weights to outcome.
- Best Model max_depth=20 and weight= 1, ntrees=50
- Metrics: Area under ROC: 0.82

  Area under PR = 0.34

- While the Area under ROC is good, Area under PR is low. This is important since the data is very imbalance.

Random forest Feature Importances



Ensemble model : Random Forest and Gradient Boosting trees.

- Performed under-sampling n data to obtain a 1:4 ratio of balance between classes.
- Trained a Random Forest Classifier - F1 score : 0.803
- Trained a Gradient Boosting Classifier : F1 score : 0.7988
- Created a stacked ensemble model with : F1 score : 0.81.
- Apply Logistic regression on predictions from Random Forest model and Gradient Boosting model.
- Ensemble model led to 17% increase in correct classification of 1s (business closings).