CONSTRUCTION PERMIT PREDICTION MODEL

JHANCY AMARSINGH

ABSTRACT:

- The main purpose of any building, shed, or electrical permits is to ensure safety. I looked at construction data from the "Washington DC Construction permit dataset," which contains information on application requesting for permit in the city of Washington DC from the year 2017 to 2021.I looked at patterns over time, trends, and correlation between variables using the matplotlib library. Encoding categorical variables, feature selection, data standardization, and scaling are performed as part of it.
- The final phase is modeling, building machine learning algorithms to predict labels. The model must predict the "ISSUED" label in order to identify the possibility of the application getting through so that the chance of application to be rejected can be found in prior without the government officials taking much efforts. I want to identity permits that are likely to be not issued in order to provide real-time feedback to submitters and/or the staff reviewing the permits.

DATA SET

https://opendata.dc.gov/datasets/construction-permits-in-2020/explore?location=38.916792%2C-77.022175%2C11.73&showTable=true

Every record identifies a construction permits information using various features. These attributes include X ,Y, OBJECTID, APPLICATIONDATE, ISEXCAVATION, ISFIXTURE, ISPAVING,ISLANDSCAPING,ISPROJECTIONS,ISPSRENTAL ,TRACKINGNUMBER , PERMITNUMBER, INTAKEDATE, ISSUEDATE , EFFECTIVEDATE , EXPIRATIONDATE, XCOORD, YCOORD , STATUS, WLFULLADDRESS, PERMITTEENAME, OWNERNAME, CONTRACTORNAME, WORKDETAIL, READYFORREVIEWDATE, APPLICANTCOMPANYNAME, LATITUDE, LONGITUDE, GIS_ID, GLOBALID, CREATOR, CREATED, EDITOR,EDITED

EDA AND DATA CLEANING:

Checked the data , info, null values ,removing the unique values , date values and imputable values.

Grouping status ‘Cancel/Withdrawn’, ‘Revise/Resubmit’, ‘Denied’, ’Suspended’, ‘Revoked’ into Not issued and Approved into Issued.

Correlation between variables , distribution of variables, check the geolocation coordinates , visualize them to see the distribution of issued and not issued.

FEATURE ENGINEERING

- One hot encoding(OHE) is a popularly used technique of categorical encoding. Here, categorical values are converted into simple numerical 1's and 0's without the loss of information.
- Rescaling of all values in a feature in the range 0 to 1

MODELING

- Independent variables are analyzed to determine the binary outcome with the results falling into one of two categories. The independent variables can be categorical or numeric, but the dependent variable is always categorical in case of classification algorithms.
- Built logistic regression model , Decision trees and support vector classifer Machine learning models with parameter tuning using grid search. Also built a sklearn neural network.

MODELING -

Logistic Regression: {'logreg__C': 100, 'logreg__penalty': 'l1', 'logreg__solver': 'liblinear'}

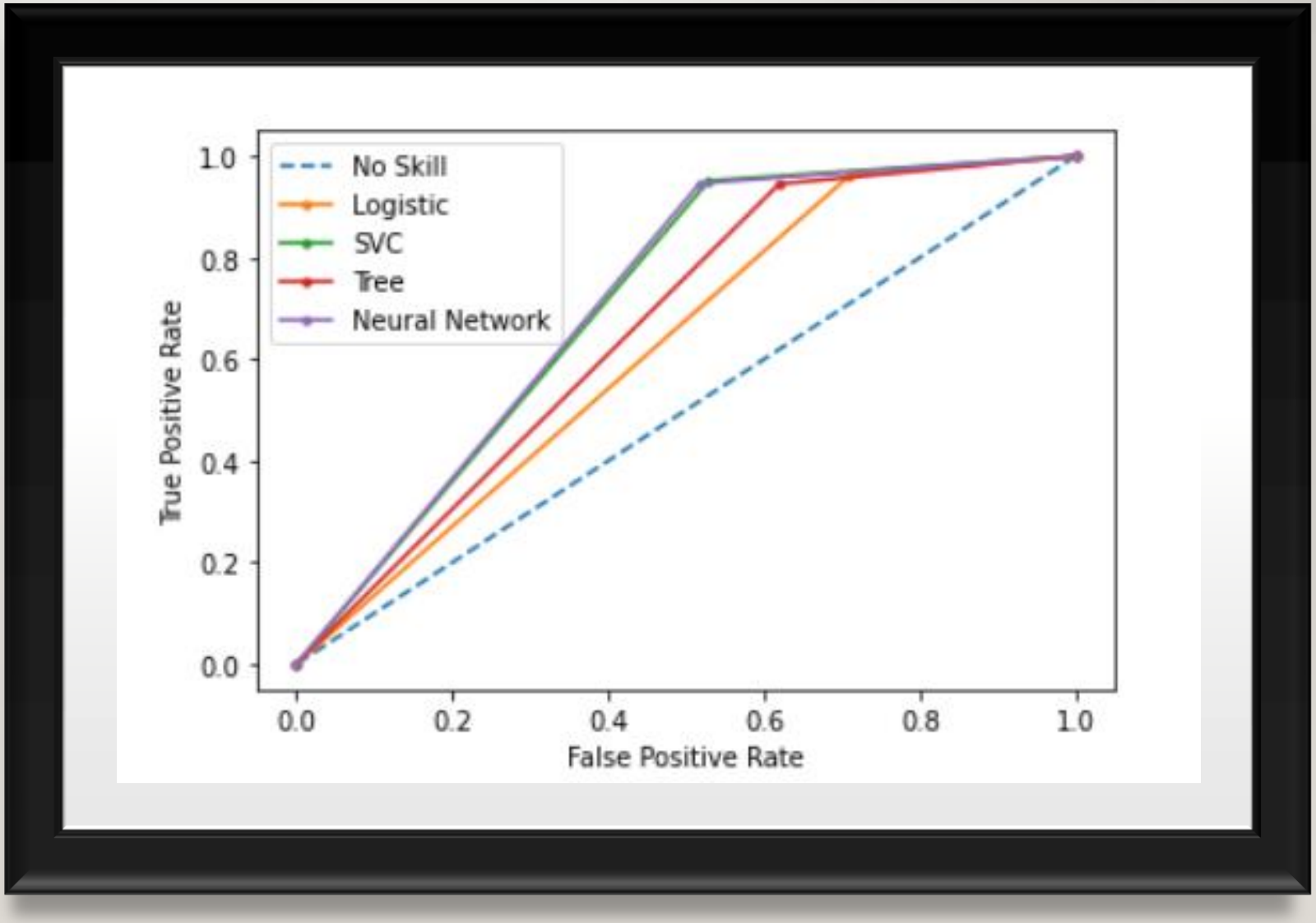
Support Vector Classifier: {'svc__kernel': 'poly'}

Decision Tree: {'tree__criterion': 'entropy', 'tree__max_depth': 10, 'tree__min_samples_leaf': 50}

	Logisitic Regression	Decision Tree	Support Vector	Neural Network
Precision	0.75	0.75	0.79	0.79
Recall	0.63	0.66	0.71	0.71
Accracy	0.83	0.83	0.85	0.85
score	0.81	0.85	0.82	0.85
roc_auc_score	0.62	0.66	0.71	0.71

RESULTS AND DISCUSSION

COMPARISON


No Skill
Logistic
SVC
Tree
Neural Network
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

FURTHER WORKS :

- More analysis with respect to four date fields in the dataset.
- Instead of just predicting two labels , (Issued and Not issued) predict more specific class.
- Extract additional variables from existing features.

VIDEO