



# Fraud Analysis

PRESENTER: DE, CARLOS B. RIVERA

COMPANY: GA ENGINEERING CONSULTANTS

AUDIENCE: DISCOVERY EXECUTIVES (NON-TECHNICAL)



# Background

- ▶ Credit card fraud is a major issue in the USA with subject matter experts saying the issue is likely to escalate due to COVID.
- ▶ Types of Credit card fraud:
  1. Online (net)
  2. Point of Sale (pos)



# Problem Statement

- ▶ This project aims to analyze credit card data for insight, and train a classifying model optimized for type 2 errors. Given the enormous amount of online transactions, an online tool will be developed on the final model to assist field partners with additional resource for classifying whether or not a given purchase is fraudulent.



# Data Information

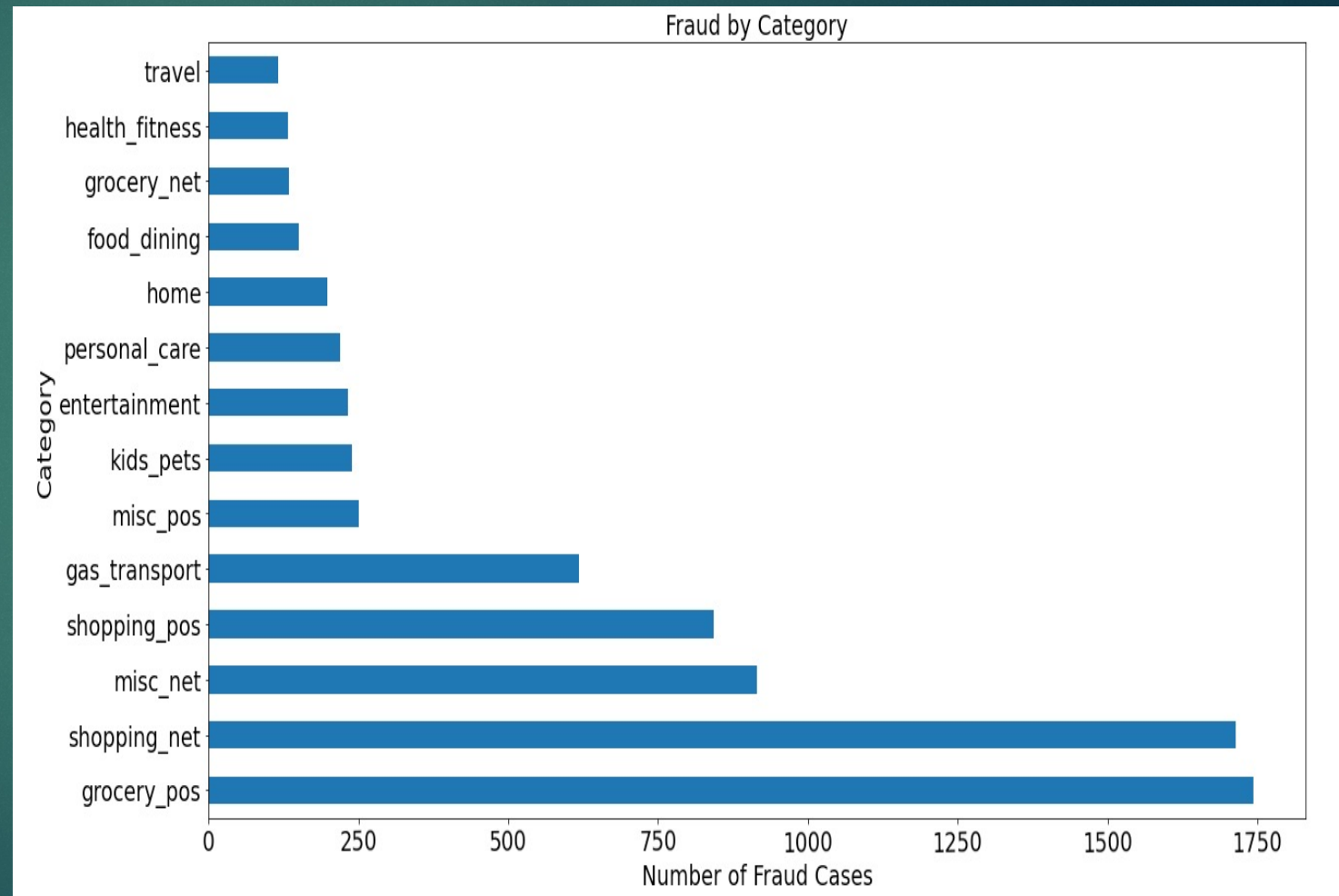
- ▶ Over 1 million cases
- ▶ Very imbalanced Classes
- ▶ EDA/Modeling performed on smaller balanced class dataset
  - ▶ 7K cases used (balanced classes)
- ▶ Modeling performed on smaller balanced class dataset
  - ▶ 350K cases used
- ▶ A New York only dataset for the online tool model
  - ▶ 85K cases used



# Initial EDA

## EDA

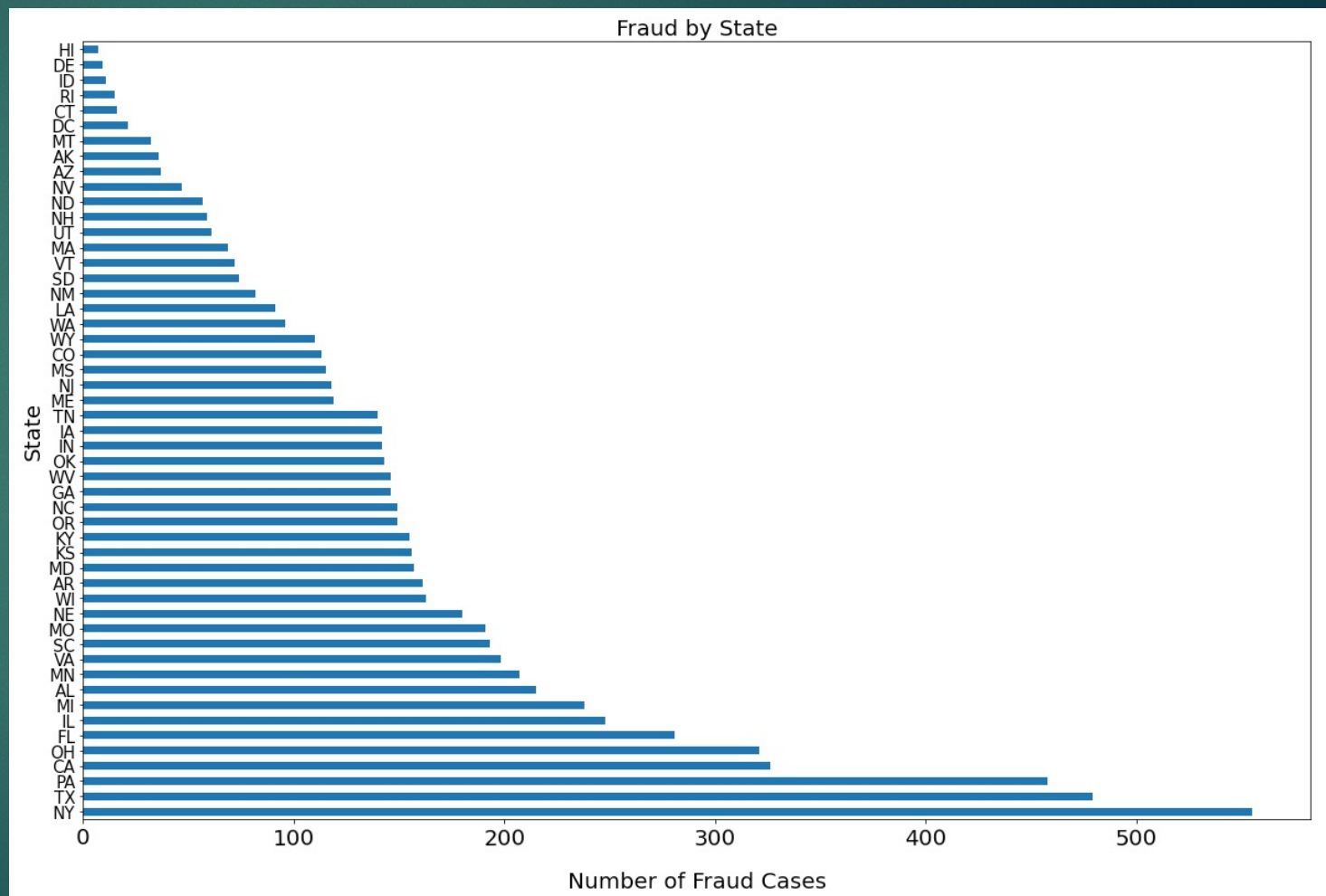
Note: Most fraudulent purchases happen online (net).





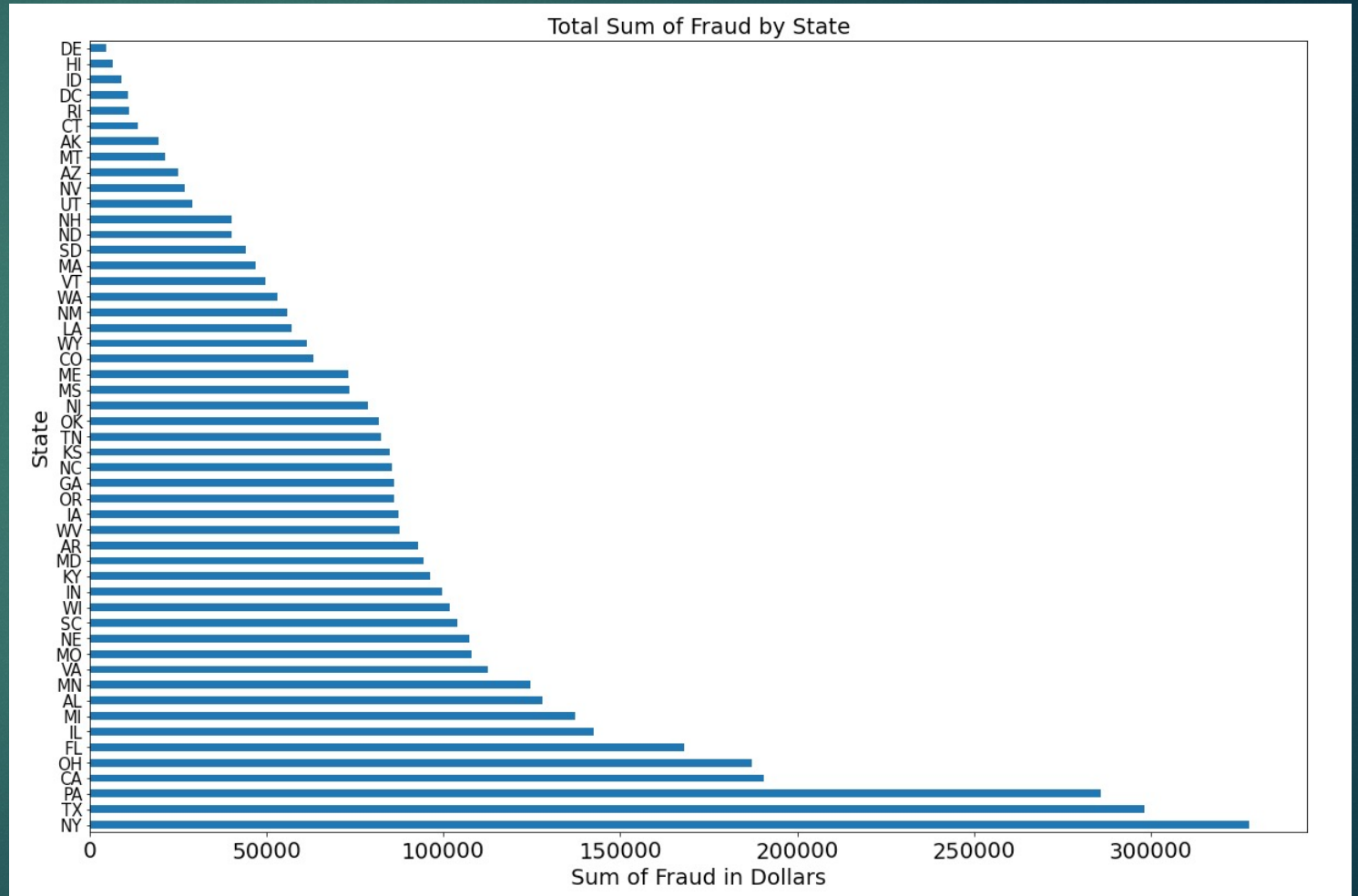
# EDA

Note: Most fraudulent cases happen in New York.



# EDA

Note: Greatest total sum of fraudulent purchases is held by New York.





# Initial Modeling

# Initial Modeling

	Train Acc	Test Acc	Acc-diff	Train-F1	Test-F1	F1-diff	Train-Pres	Test-Pres	Pres-diff	Train_Recall	Test-Recall	Recall_diff
<b>knn</b>	0.765166	0.667732	0.097433	0.785285	0.696815	0.088470	0.723369	0.640590	0.082779	0.858792	0.763859	0.094933
<b>logreg</b>	0.899547	0.839329	0.060218	0.898592	0.839071	0.059521	0.907297	0.840192	0.067104	0.890053	0.837953	0.052100
<b>dt</b>	1.000000	0.963229	0.036771	1.000000	0.963569	0.036431	1.000000	0.954498	0.045502	1.000000	0.972814	0.027186
<b>bag</b>	0.995470	0.963496	0.031974	0.995464	0.963785	0.031678	0.996971	0.955952	0.041020	0.993961	0.971748	0.022213
<b>bag_knn</b>	0.776090	0.670397	0.105693	0.793614	0.696888	0.096726	0.736067	0.644898	0.091169	0.860924	0.757996	0.102928
<b>bag_log</b>	0.901146	0.836930	0.064215	0.899685	0.835307	0.064378	0.913266	0.843478	0.069788	0.886501	0.827292	0.059209
<b>rf</b>	1.000000	0.966160	0.033840	1.000000	0.966124	0.033876	1.000000	0.966898	0.033102	1.000000	0.965352	0.034648
<b>et</b>	1.000000	0.938449	0.061551	1.000000	0.937347	0.062653	1.000000	0.954169	0.045831	1.000000	0.921109	0.078891
<b>ada</b>	0.947509	0.944578	0.002931	0.947019	0.944116	0.002903	0.956018	0.951788	0.004230	0.938188	0.936567	0.001621
<b>gboost</b>	0.965006	0.961364	0.003642	0.964890	0.961385	0.003506	0.968169	0.960617	0.007551	0.961634	0.962154	0.000519
<b>svc</b>	0.905675	0.820677	0.084999	0.902640	0.808970	0.093671	0.932740	0.865209	0.067530	0.874423	0.759595	0.114828

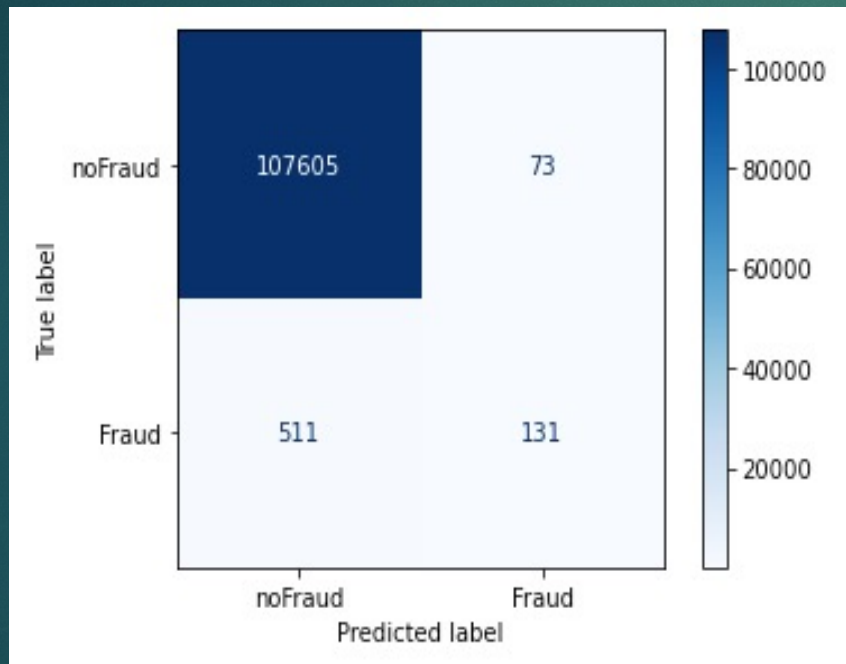


# Initial Modeling

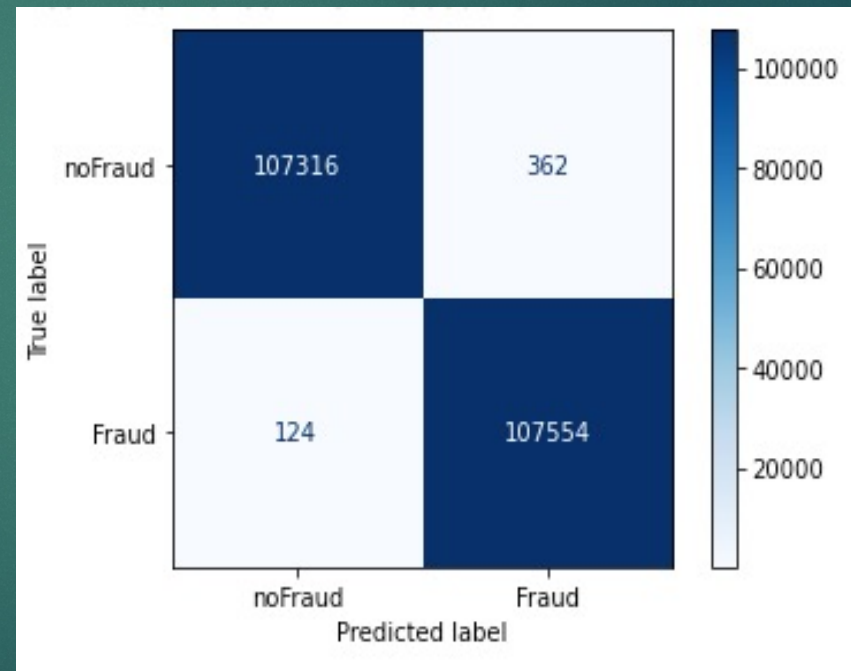
	Train Acc		Test Acc	Acc-diff	Train-F1		Test-F1	F1-diff	Train-Pres		Test-Pres	Pres-diff	Train_Recall	Test-Recall	Recall_diff
<b>dt</b>	1.0	0.963229	0.036771		1.0	0.963569	0.036431		1.0	0.954498	0.045502		1.0	0.972814	0.027186
<b>rf</b>	1.0	0.966160	0.033840		1.0	0.966124	0.033876		1.0	0.966898	0.033102		1.0	0.965352	0.034648
<b>et</b>	1.0	0.938449	0.061551		1.0	0.937347	0.062653		1.0	0.954169	0.045831		1.0	0.921109	0.078891

# After Hyper-Tuning

Null Model(Logistic Regression - default)



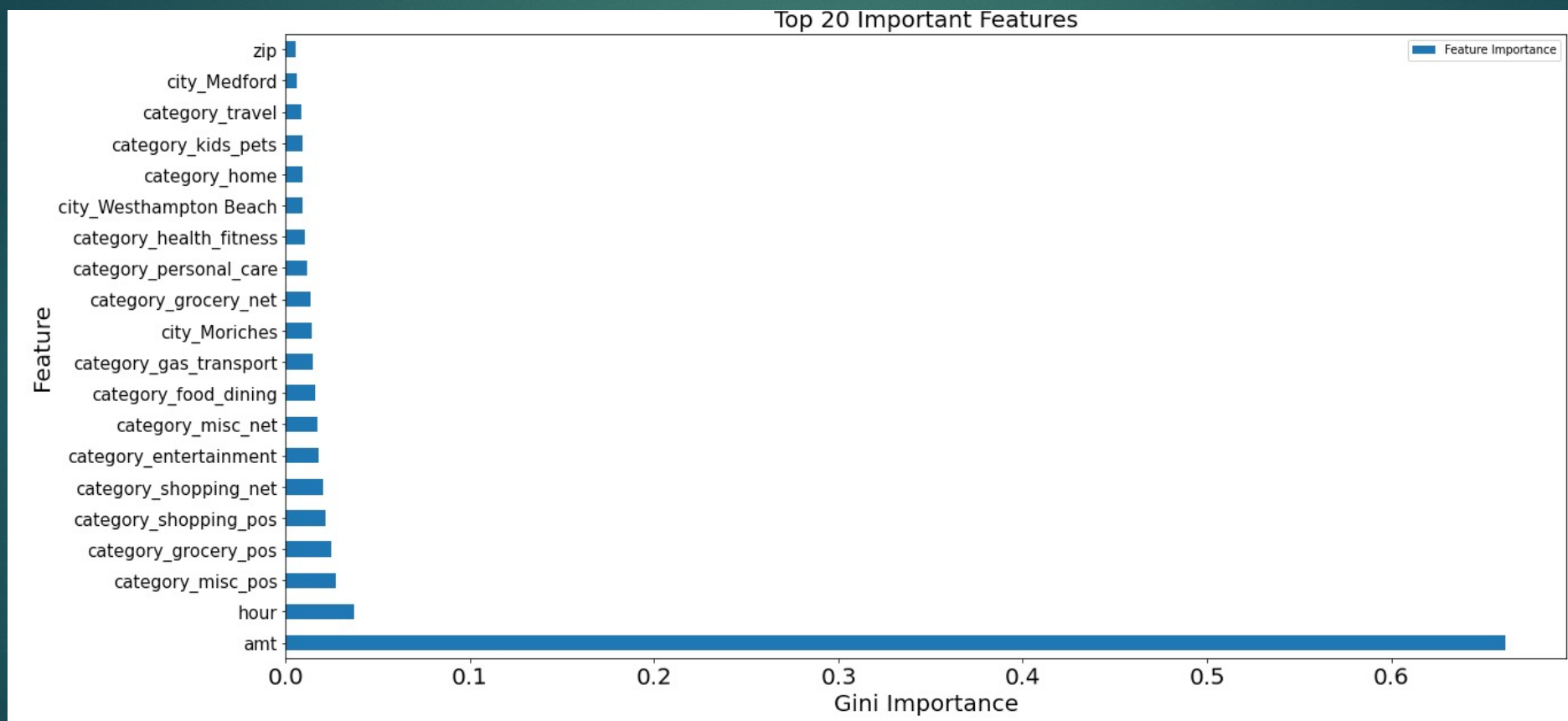
Decision Tree Model (default settings)



Note: Both Models had high accuracy with very little overfitting.



# Results EDA

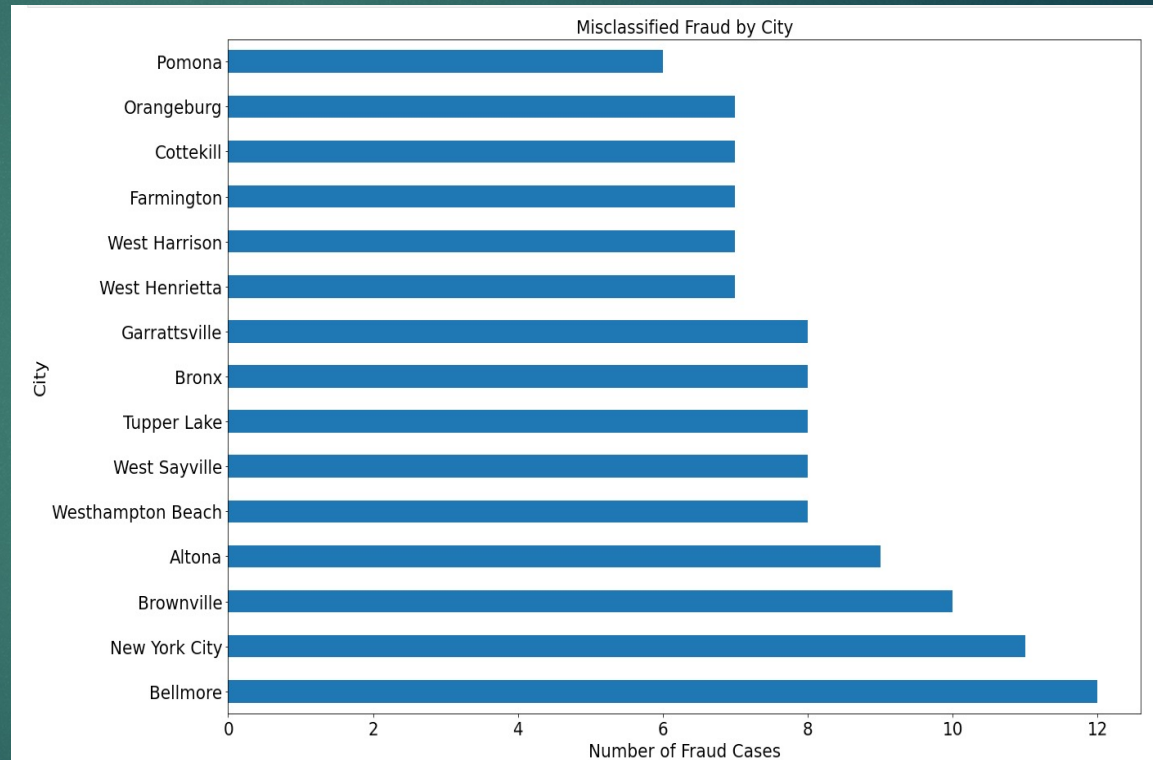


# Misclassification EDA



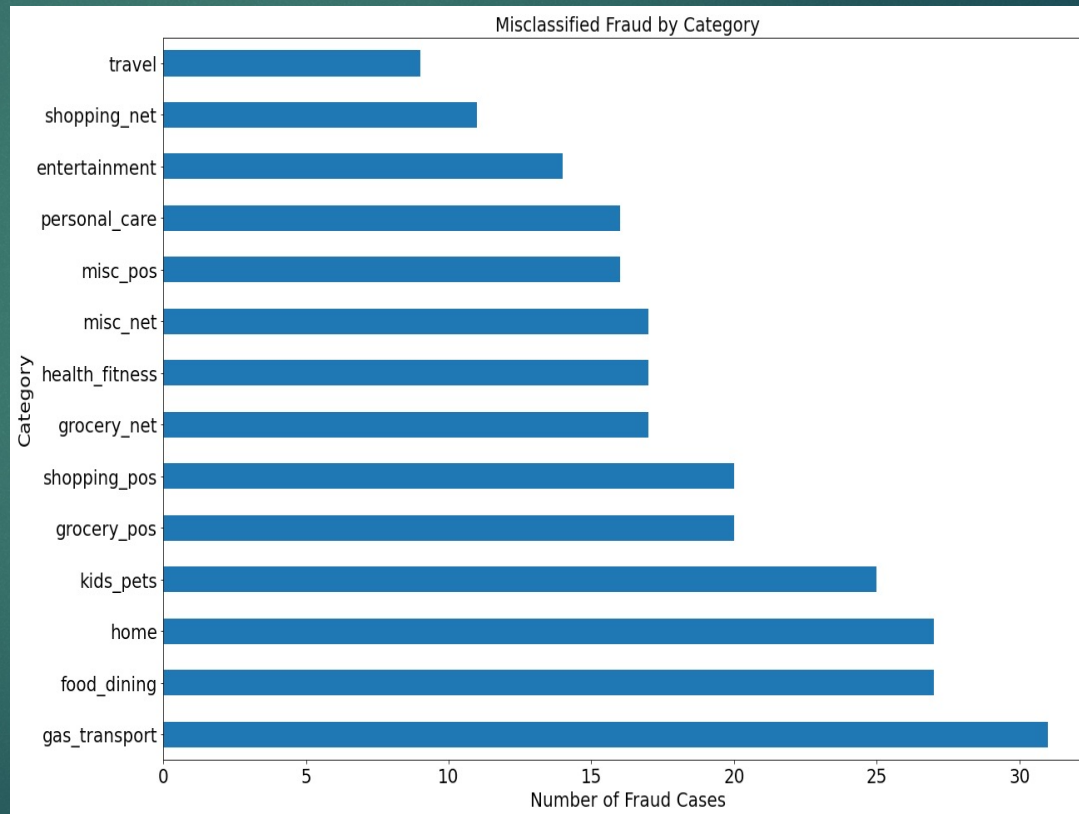
# Misclassified EDA

city	
Brooklyn	[2504700]
New York City	[1577385]
Bronx	[1382480]
Albany	[151022]
South Richmond Hill	[47211]
North Tonawanda	[45100]
Bellmore	[34496]
Northport	[21902]
Hudson	[17867]
West Harrison	[11250]



# Misclassification EDA

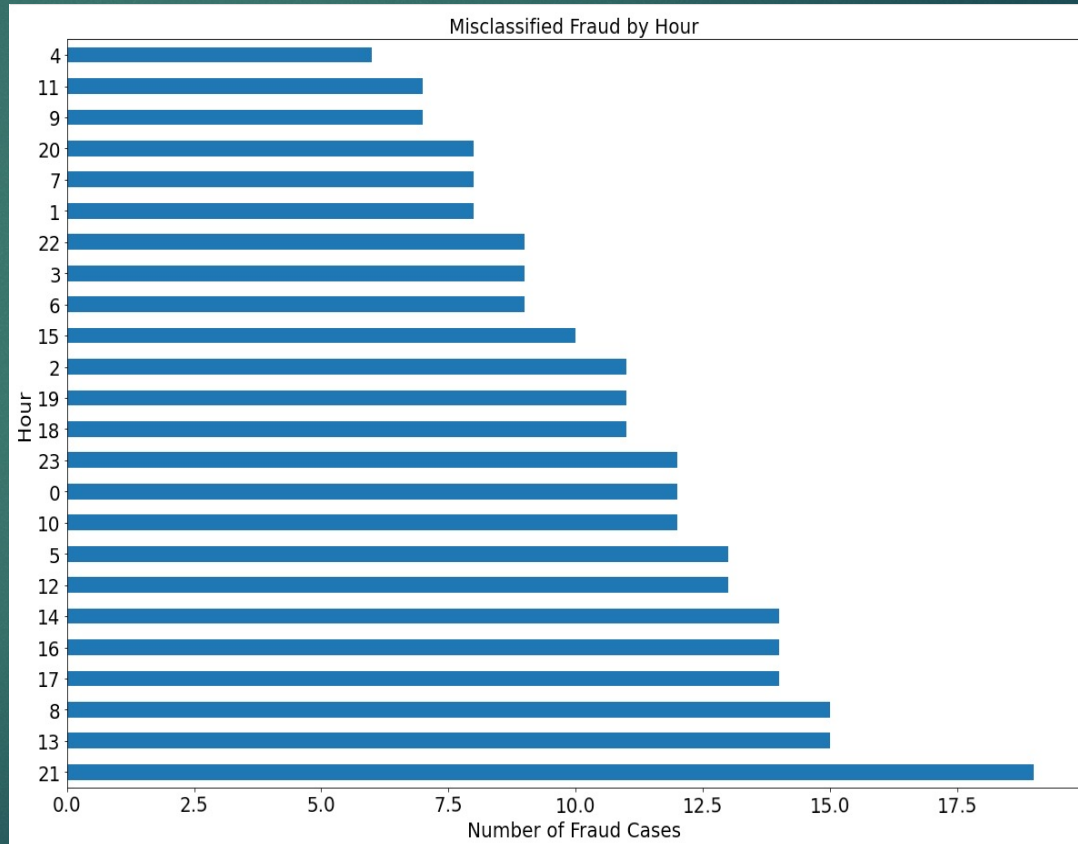
Note: Most misclassified fraud cases in NY happen while the card is in hand.





# Misclassification EDA

Note: Most misclassified fraud purchases in New York happens during the 9pm.



# Fraud App Demo

[Go!](#)



# Conclusion

- ▶ Although my models aren't perfect, they classify very well with very little overfitting.
- ▶ Our Final model significantly decreased type 2 errors.



# Field Partner Recommendations

- ▶ Misclassifications that happen after late hours are not always fraud.
  - ▶ The app is designed to only offer input, so no matter the classification double check with customer.