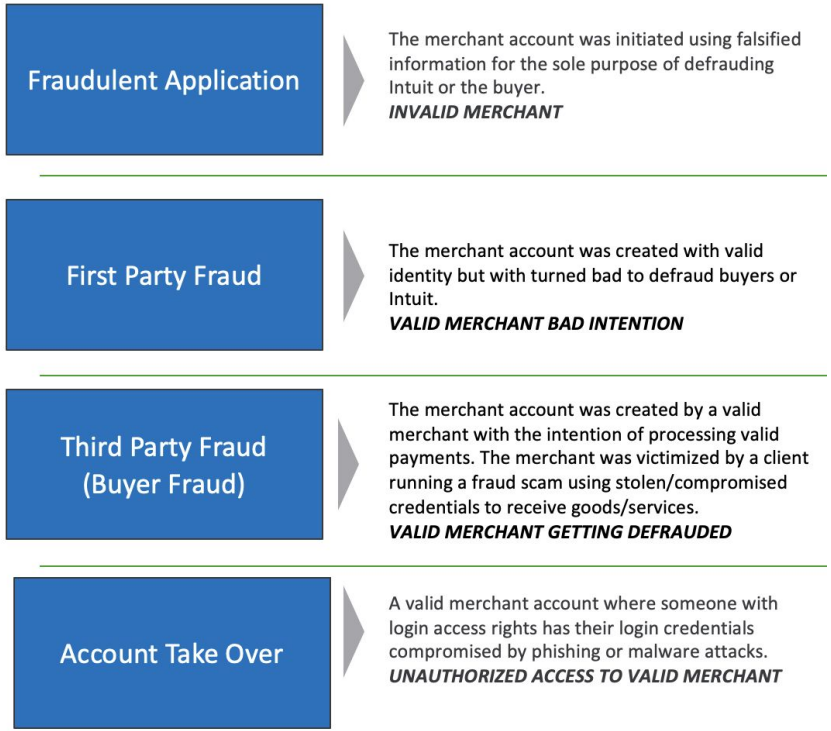
Classification of QBO Applications for Fraud Risk

Shubham Sah (Intuit DS Cohort 8)

April 2021

Project Outline

- Merchants apply for Intuit QBO Payments and provide their personal details (name, address, business details, DOB, SSN etc)
- Application data enriched using features from several external vendors such as Experian (credit risk), LexisNexis. White Pages and Email Age (fraud risk), Giact (Bank Evaluation)
- The model needs to predict whether an application will turn out to be fraudulent.

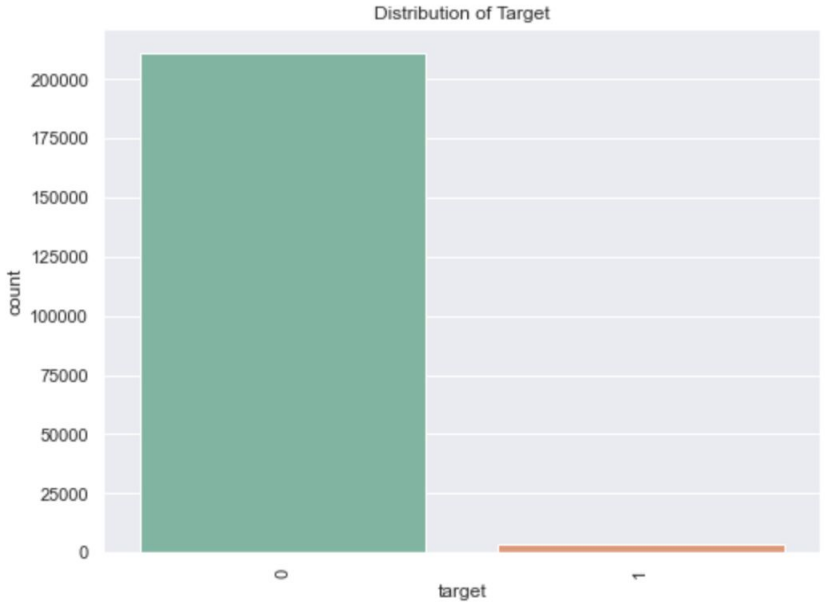


Data (and Preprocessing)

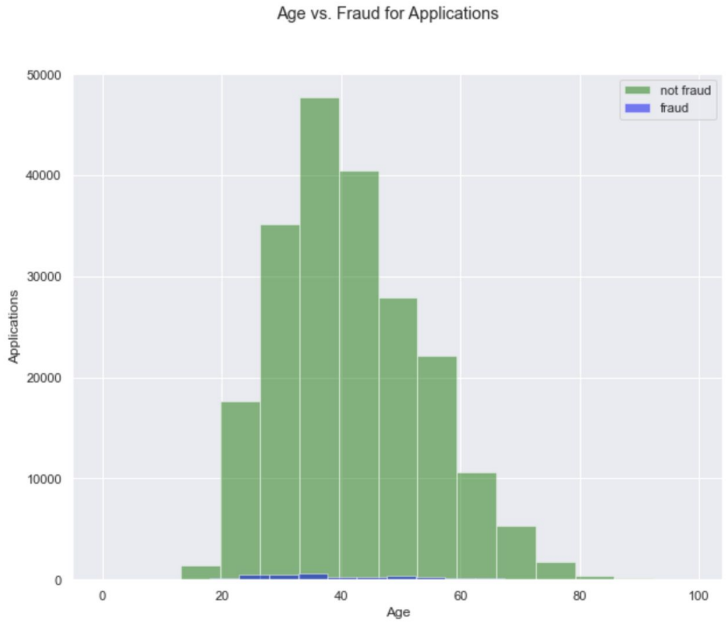
- Payment Applications from July 2020 to December 2020
- ~220k observations
- 65 Features (44 numerical, 3 dates and 18 categories)

- Converted dates to days difference from application date
- Replaced missings by Means, Medians, NAs etc
- Dropped few irrelevant features (unique ids, dates etc)

Sample EDA

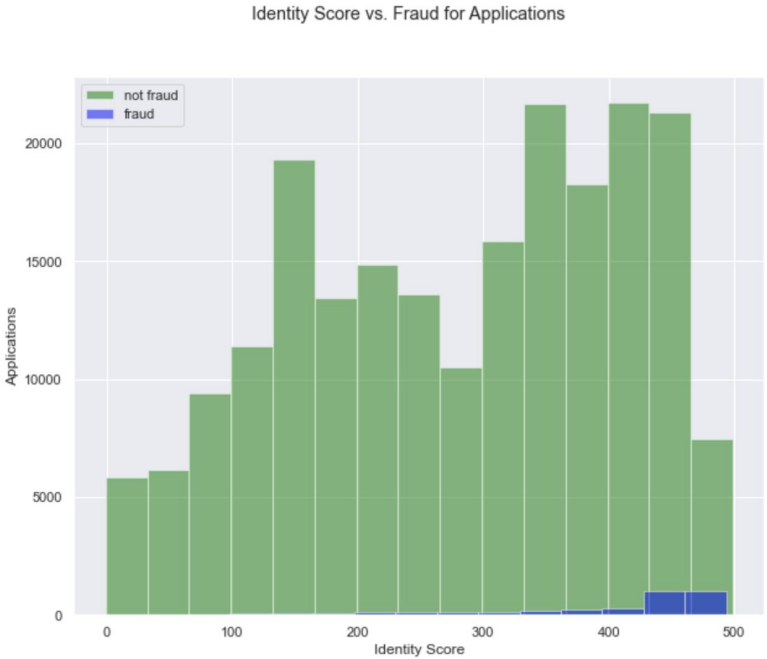


Target Class Highly Imbalanced



Most applicants between 35 and 45 years of age

Sample EDA



Most Fraud can be targeted by score ranges bw 400 and 500

Most Fraud can be targeted by scores =500 and >900

Feature Selection, Scaling and Encoding

- Using Seaborn Pairplot and Heatmap, checked the relationship between a few hand picked attributes
- Dropped more features which were not adding much value
- Scaled the continuous features using Standard Scaler
- Converted all categorical features to numbers using One Hot Encoder

Model Selection and Tuning

- After all cleaning and transformations, we have ~214k observations with 114 features
- Split in training and testing sets
- Use several classification techniques for model selection
- Random Forest Classifier seems to work the best
- Due to the highly imbalanced data, model accuracy is disregarded as a metric
- ROC Curves, AUC, Precision and Recall are selected as the validation metrics

Conclusion - Model Comparison

Model Selection

Model Tuning

	Accuracy	AUC	Precision	Recall	F1 Score
KNN Model	0.987	0.80			
Logistic Regression	0.985	0.89			
LinearSVC	0.984	0.88			
Decision Tree Classifier	0.980	0.74			
Gradient Boosting Classifier	0.987	0.92			
Random Forest	**0.990**	**0.93**	**0.99**	**0.42**	**0.59**
Random Forest with Hyper Parameter Tuning		0.94	1.00	0.27	0.42
Adjusting Probability Thresholds (0.5→0.2)		0.94s	0.76	0.41	0.54
Undersampling		0.93	**0.87**	**0.88**	**0.87**

Next Steps

- More rows and columns (larger slice and more features)
- More Vendors and larger ecosystem data (scope)
- Oversampling and Synthetic Data
- Use Modern ML Techniques (Neural Networks/Deep Learning)