Fraud Detection from a Machine Learning Perspective

Filipa Peleja
Lead Data Scientist at Levi Strauss & Co Europe
27 May 2021

Fraud Detection

"Fraud detection is a set of activities that are taken to prevent money or property from being obtained through false pretences"

Rank	Category	# of Reports
1	Internet Services	62,942
2	Credit Cards	51,129
3	Healthcare	47,410
4	Television and Electronic Media	38,336
5	Foreign Money Offers and Counterfeit Check Scams	27,443
6	Computer Equipment and Software	18,350
7	Investment-Related	14,884

https://spd.group/machine-learning/credit-card-fraud-detection/

Filipa Peleja – filipapeleja@gmail.com

Let’s think about
Credit Card Fraud Detection

Filipa Peleja – filipapeleja@gmail.com

Machine Learning Card Fraud Detection

- Detect fraud automatically
- Real-time streaming
- Less time required for methods validation
- Identify hidden correlations in data

Conventional Fraud Detection

- Decision making for fraudulent schemes should be set manually
- Time consuming
- Multiple verification methods are needed; thus, not ideal for the user
- Finds only obvious fraud activities

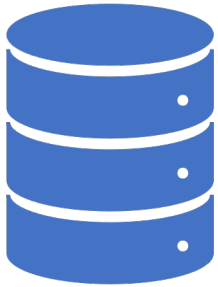
https://spd.group/machine-learning/credit-card-fraud-detection/

Filipa Peleja – filipapeleja@gmail.com

Machine Learning for Fraud Detection – **How?**

Filipa Peleja – filipapeleja@gmail.com

What does Machine Learning Models Need?



Data availability and volume

Quality of data

Filipa Peleja – filipapeleja@gmail.com

Steps taken in a Machine Learning pipeline for Fraud Detection

01 Data Exploration

For data understanding, preparation, grouping and segmenting data from millions of transactions to find patterns and detect fraud

02 Pattern Recognition

Machine Learning offers a choice from a set of models that best fit certain business problems

03 Predictive Models

Learn from prior data and estimate the probability of a fraudulent credit card transaction

Filipa Peleja – filipapeleja@gmail.com

Unsupervised models

Machine Learning algorithms that learn patterns from unlabelled data

Objective is to model the underlying structure or distribution in the data so we can learn more from it

For unsupervised models, there are no “correct answers”(labels) and there’s no “teacher”

Unsupervised models are left to their own devises to discover and present the interesting structure in the data

https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Filipa Peleja – filipapeleja@gmail.com

Unsupervised models for Fraud Detection

- **Principal Components Analysis**
  - Enables to reveal the inner structure of the data and explain it's variations
  - One of the most popular techniques for Anomaly Detection

- **One-class Support Vector Machine[1]**
  - Fit a model on the "normal" data and predict whether new data is normal or an outlier/anomaly

- **Isolation Forest**
  - Anomaly detection from the family of decision trees algorithms
  - Detects anomalies instead of profiling the positive data points

[1] https://machinelearningmastery.com/one-class-classification-algorithms/

Supervised models

Learn from historical data

Algorithms are trained using labelled data

Objective is to train from a data that represents well the population and can generalise it's learnings well in un-seen data

Can be categorized in Classification and Regression problems

Filipa Peleja – filipapeleja@gmail.com

Supervised models for Fraud Detection

- **K-Nearest Neighbours**
  - Finds similarities based in distance in multi-dimensional space
  - Simple model, not vulnerable to noise or missing data points

- **XGBoost (Extreme Gradient Boosting) and Light GBM (Gradient Boosting Machine)**
  - Single type of gradient-boosted decision trees algorithms which are more effective in computing time and memory resources
  - This algorithm is a blending technique where new models are added to fix the errors cause by previous models

- **Random Forest**
  - Uses ensemble technique that uses many decision trees
  - The random and diverse nature of random forest (rows/columns are randomly chose) helps to accommodate the finding for new fraud schemes

More…

- There are more machine learning models that can help Fraud Detection e.g., Neural Networks and simple models like Decision Trees that can be easily interpreted and visualised

- Fraud detection models should always be continuously learning as new data arrives to capture schemas/patterns as early as possible

Exploration for Insurance Fraudulent Claims

Filipa Peleja – filipapeleja@gmail.com

My data has 39 features/columns

months_as_customer	age	policy_number	policy_bind_date	policy_state	policy_csl	policy_deductable	policy_annual_premium	umbrella_limit	insured_zip	insured_sex
328	48	521585	17-10-2014	OH	250/500	1000	1406.91	0	466132	MALE
228	42	342868	27-06-2006	IN	250/500	2000	1197.22	5000000	468176	MALE

insured_education_level	insured_occupation	insured_hobbies	insured_relationship	capital-gains	capital-loss	incident_date	incident_type	collision_type	incident_severity	authorities_contacted	incident_state	incident_city
MD	craft-repair	sleeping	husband	53300	0	25-01-2015	Single Vehicle Collision	Side Collision	Major Damage	Police	SC	Columbus
MD	machine-op-inspct	reading	other-relative	0	0	21-01-2015	Vehicle Theft	?	Minor Damage	Police	VA	Riverwood

incident_location	incident_hour_of_the_day	number_of_vehicles_involved	property_damage	bodily_injuries	witnesses	police_report_available	total_claim_amount	injury_claim	property_claim	vehicle_claim	auto_make	auto_model	auto_year	fraud_reported
9935 4th Drive	5	1	YES	1	2	YES	71610	6510	13020	52080	Saab	92x	2004	Y
6608 MLK Hwy	8	1	?	0	0	?	5070	780	780	3510	Mercedes	E400	2007	Y

Filipa Peleja – filipapeleja@gmail.com

My data – Target Variable is **Fraud Reported**

```
insurance_df['fraud_reported'].value_counts()
```

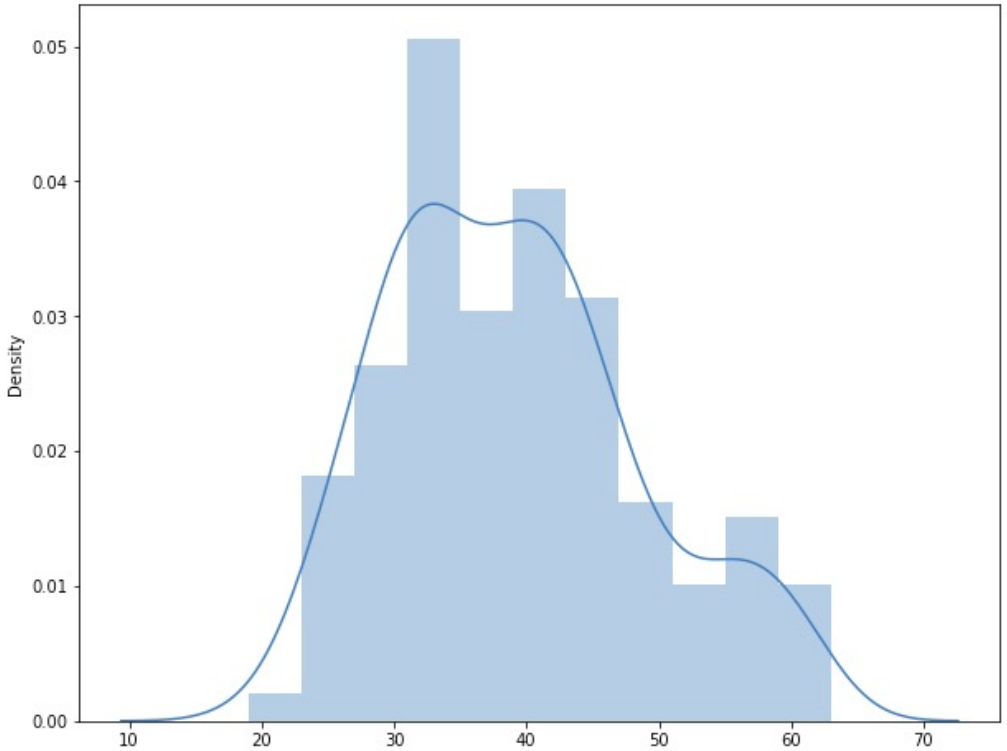
```
N    753
Y    247
```

Imbalanced problem

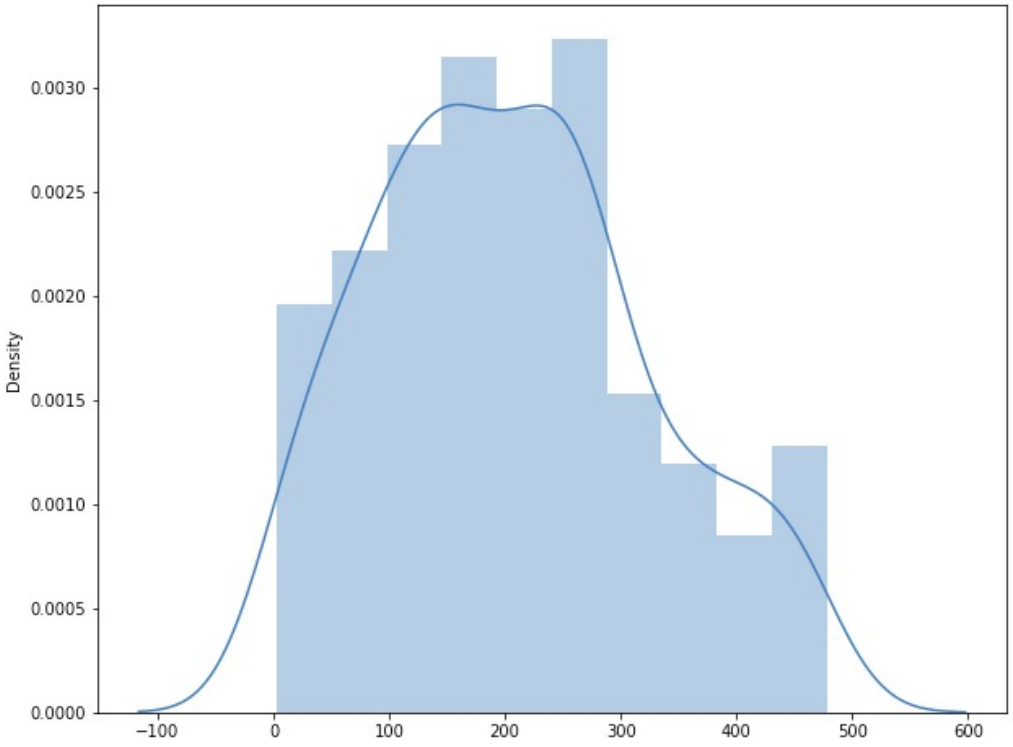
injury_claim	property_claim	vehicle_claim	auto_make	auto_model	auto_year	fraud_reported
6510	13020	52080	Saab	92x	2004	Y
780	780	3510	Mercedes	E400	2007	Y

Filipa Peleja – filipapeleja@gmail.com

Let’s observe the Fraud Reported distribution for Age and Months as Customer



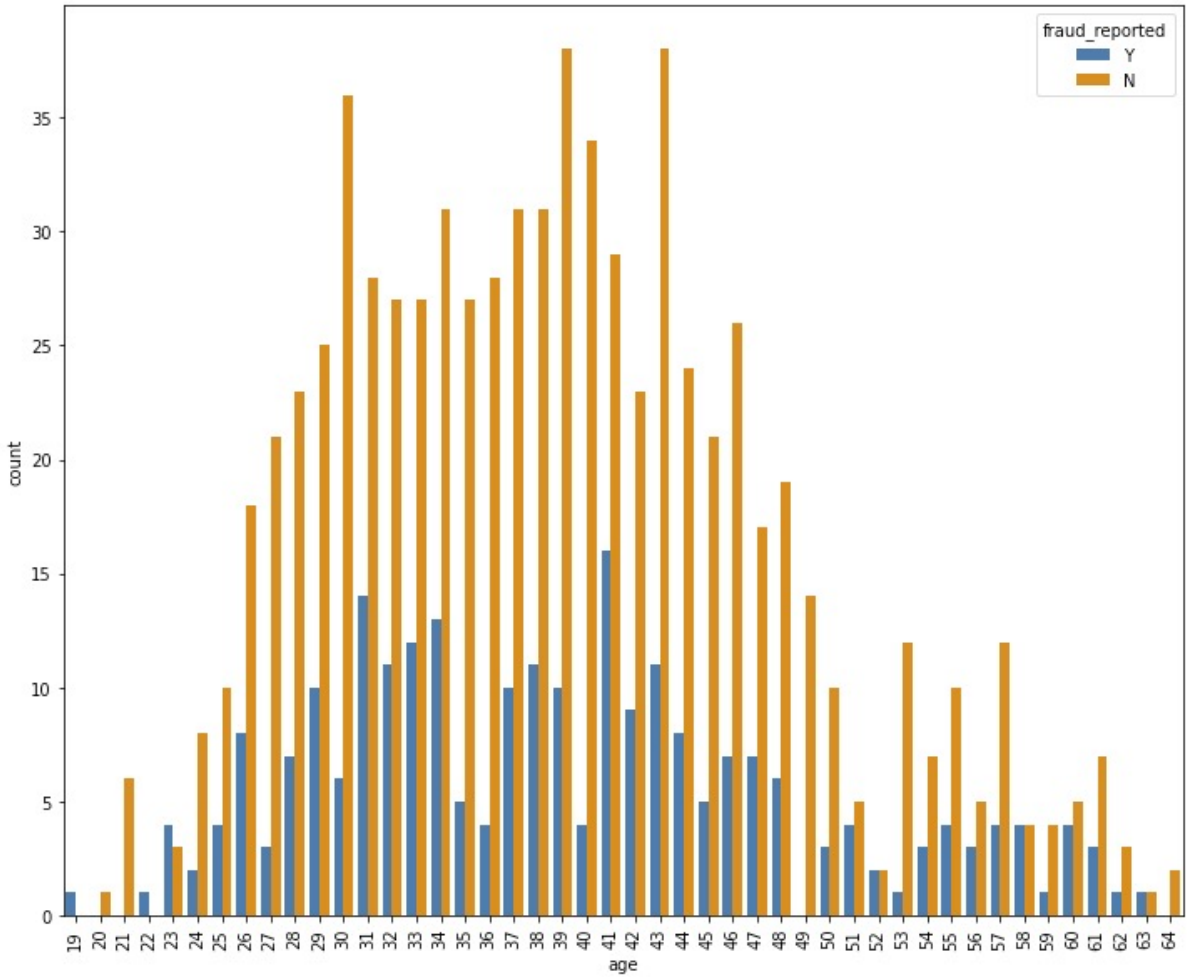
Age



Months as Customer

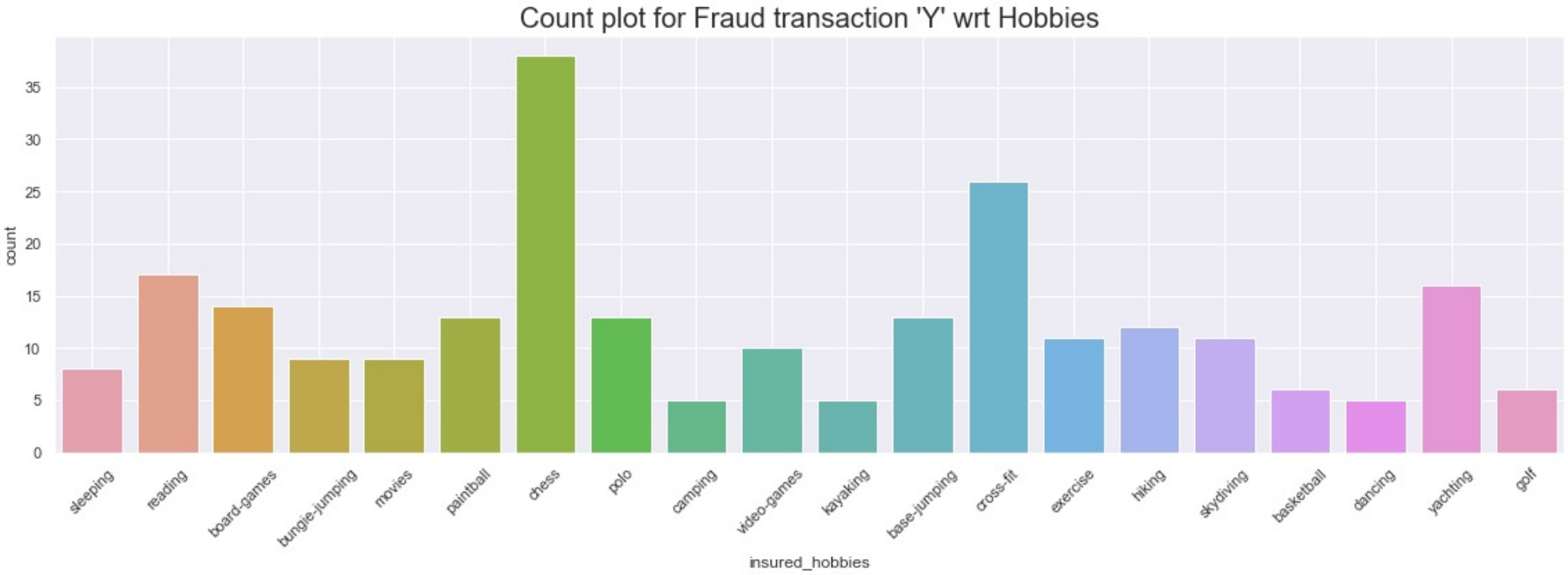
Filipa Peleja – filipapeleja@gmail.com

Is there a correlation between Age and Fraud Reported?



Filipa Peleja – filipapeleja@gmail.com

What about Customer hobbies?



Filipa Peleja – filipapeleja@gmail.com

More and more exploration to better understand the data and relevant Features for the model

Filipa Peleja – filipapeleja@gmail.com

This is a Supervised Learning Approach

The selected Machine Learning algorithm is Random Forest

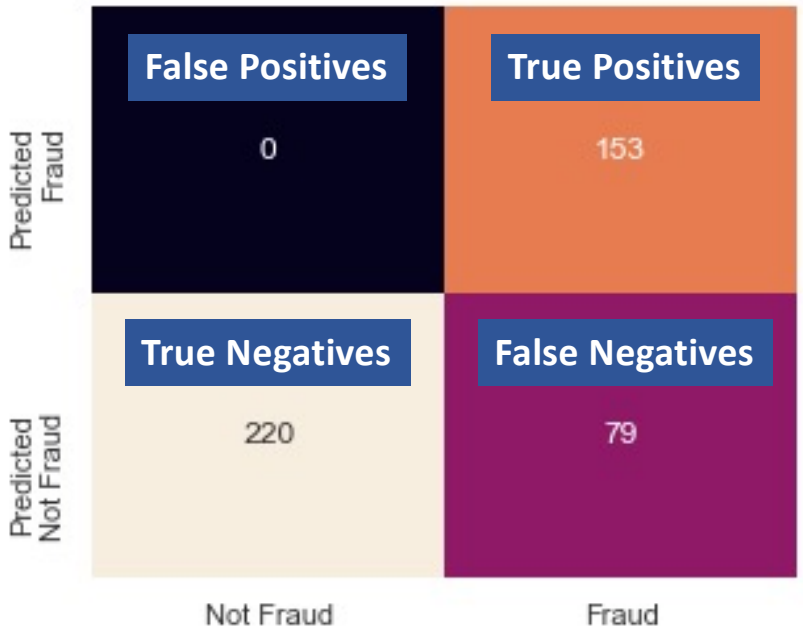
```
# Build a RF model
# Takes some time to compute ~5minutes
x_train, x_test, y_train, y_test = \
train_test_split(
    x_scale,
    y_upsample,
    test_size=0.3
)
```

```
# Now we know the best estimator and select that model
rf = RandomForestClassifier()

parameters = {
    'criterion':[
        'gini',
        'entropy'
    ],
    'max_depth': np.arange(1,30)
}

grid = GridSearchCV(rf, parameters)


# Find the best parameters for my train data
grid.fit(x_train, y_train)
```



Filipa Peleja – filipapeleja@gmail.com

Thank you!

Filipa Peleja

filipapeleja@gmail.com