Ensemble Methods and Random Forests

Jupyter notebooks

- Decision Trees
- Random Forests
- Gradient Tree Boosting

🔗 Homework ()

- Credit scoring problem
- Flight delays
- Forest Cover Type Prediction

In-class project ()

- Credit Card Fraud Detection

Datasets

Filename	Description	Source
good_or_evil_train good_or_evil_test	Characters in the Batman universe	
flights2015	national flights that took place in the US territory in 2015 as well as information about the airlines and the airports	U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics, also Kaggle
credit_scoring	Data on credit scoring	Kaggle

Filename	Description	Source
forest	The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m x 30m patch	US Forest Service (USFS) Region 2 Resource Information System data aslo Kaggle

Project Dataset

The **Data folder** contains the following data:

- Untagged_Transactions.csv: transaction data from on an online store
- Account_Info.csv: Anonomyzed account information
- Fraud_Transactions.csv: transactions identified as fraud

Untagged_Transactions.csv contains the following fields:

Columns	Description
transactionID	Unique transaction Id
accountID	Unique account Id
transactionAmountUSD	Transaction amount in USD e.g., 12345.00
transactionAmount	Transaction amount in currency expressed in transactionCurrencyCode e.g., 12345.00
transactionCurrencyCode	Currency code of the transaction. 3 alphabet letters, e.g., USD
transactionCurrencyConversionRate	Conversion rate to US Dollars, e.g. 1.0000 for USD to USD
responseCode	response code from card issuer payment authorization
digitalItemCount	Number of digital items purchased. (e.g. music, ebook, software, etc, that can be directly downloaded online)

Columns	Description
physicalItemCount	Number of physical items purchased (that needs to be shipped)
purchaseProductType	Type of product purchased
shippingAddress	shipping street address
shippingPostalCode	shipping postal code
shippingCity	shipping city
shippingState	shipping state
shippingCountry	shipping country (3-alpha)
cvvVerifyResult	M-- CVV2 Match N-- CVV2 No Match P--Not Processed S--Issuer indicates that CVV2 data should be present on the card, but the merchant has indicated data is not present on the card U--Issuer has not certified for CVV2 or Issuer has not provided Visa with the CVV2 encryption keys Empty--Transaction failed because wrong CVV2 number was entered or no CVV2 number was entered
paymentInstrumentID	ID of payment Instrument: e.g. credit card number (hashed or encrypted) e.g. paypal account Id
paymentBillingAddress	Street Address , hashed or encrypted
paymentBillingPostalCode	payment billing postal code
paymentBillingState	payment billing state
paymentBillingCountryCode	payment billing country code
paymentBillingName	Name, hashed or encrypted, needs to be consistent with other names
isProxyIP	Whether the IP address is a proxy or not

Columns	Description
browserType	I -- IE C -- Chrome F -- Firefox O -- Other
browserLanguage	Similar to country code
paymentInstrumentType	Type of payments: C -- Credit Card D -- Debit Card P -- Paypal K -- Check H -- Cash O -- Other
cardType	Type of cards M -- Magnetic C -- Chip
cardNumberInputMethod	Input method of payment instrument number: K -- Keyed S -- Swiped C --- Chip D -- Contactless
transactionDeviceType	P -- PC M -- Mobile Devices C -- Console (e.g. Xbox, DVD) O -- Other
transactionDeviceId	Mac Address, or Hardware ID like serial number
transactionIPaddress	Full IP Address for IPv4: 000.000.000.000
ipState	State of IP address originated from 2 alphabet letters
ipPostcode	Postal Code of IP address originated from
ipCountryCode	Country code of IP address originated from

Columns	Description
transactionDate	Date when transaction occured Typically in the time zone of the processor, Format: yyyymmdd, e.g., 20000101
transactionTime	Time when transaction occurred. Typically in the time zone of processing end. Format: hhmmss, eg. 153059
localHour	The hour in local time. Value of 0-23
transactionScenario	A -- Authorization O -- Others
transactionType	Type of transaction: P -- Purchase R -- Refund T -- Transfer O -- Other
transactionMethod	I -- Internet (Online) Order P -- Phone order M -- Mail order O -- Other

Account_Info.csv contains the following fields:

Columns	Description
transactionDate	Date when transaction occured Typically in the time zone of the processor. Format: yyyymmdd, e.g., 20000101
transactionTime	Time when transaction occurred. Typically in the time zone of processing end. Format: hhmmss, eg. 153059
accountOwnerName	User name (hashed/encrypted)
accountAddress	User street address
accountPostalCode	User postal code
paymentInstrumentAgeInAccount	Age of payment instrument in the account

Columns	Description
numPaymentRejects1dPerUser	Number of payment rejection in one day of this user
accountCity	User city
accountState	User state
accountCountry	User country (3-alpha)
accountOpenDate	Account open date. Format: yyyymmdd
accountAge	Age of user account in number of days
isUserRegistered	Whether the user is registered or not

Fraud_Transactions.csv contains the following fields:

Columns	Description
transactionID	Unique transaction Id
accountID	Unique account Id
transactionAmount	Transaction amount in currency expressed in transactionCurrencyCode e.g., 12345.00
transactionCurrencyCode	Currency code of the transaction. 3 alphabet letters, e.g., USD
transactionDate	Date when transaction occured Typically in the time zone of the processor. Format: yyyymmdd, e.g., 20000101
transactionTime	Time when transaction occurred. Typically in the time zone of processing end. Format: hhmmss, eg. 153059
localHour	The hour in local time. Value of 0-23
transactionDeviceId	Mac Address, or Hardware ID like serial number
transactionIPaddress	Full IP Address for IPv4: 000.000.000.000