2ND ROUND INTERVIEW FOR DS POSITION IN TELESIGN: DESCRIPTION OF THE RESEARCH TASKS

1. Description of the Project

Fraud in financial sector, telecom industry, on the Internet, and many other industries became a big business. Phone calls, credit card numbers and stolen accounts can be abused with a significant profit. Large profits, expansion of information technology and modern ways of communication caused the growth of a well-organized and well-informed community of fraudsters, causing huge financial losses every year all over the world.

There are two common methods used to reduce fraud:

- Fraud prevention systems - represents a set of measures to prevent the realization of fraud, such as: PIN for bank cards, SIM cards for cell phones, passwords for computer systems, one-time password with two factor authentication, etc.
- Fraud detection systems - represents a set of mechanisms to identify fraud in the shortest possible time interval from the moment it was committed.

Types of fraud TeleSign dealing with have certain characteristics that make them particularly attractive to fraudsters:

- danger of localization is small,
- no particularly sophisticated equipment is needed,
- the product of fraud could be directly convertible to money.

In this research task we want to explore the impact of IP behavior on the fraud detection systems, i.e. to explore how we can detect bad end-users based on the information about their IP addresses. We have two different data sources for the IP data, which offer the IP attributes that correspond to each end-user IP address, which was seen on the TeleSign platform during the account creation process. Each data field from the sample, corresponding to a particular data source, has the corresponding suffix *_source1* or *_source2*. Data are presented in the file *TS_Sample_DS_candidates.csv* (∼48K rows/transactions, but not necessarily distinct IP addresses).

Column *fraud_label* is our target variable, which contains information regarding the confirmed fraudulent or legitimate behavior of an end-user.

The main goals of this task would be:

(1) to analyze the quality of the IP data offered/presented in the sample *TS_Sample_DS_candidates.csv*; to perform comparative analysis of the two data sources in order to make conclusions about the accuracy of the data, i.e. to make the decision (if possible) which data source is the most accurate and recommended for the further research,
(2) to conduct the research regarding the predictive power of the IP data (from the sample mentioned above) for the fraud detection problem (i.e. predictive power regarding the target variable *fraud_label*),
(3) to try to identify specific patterns in the end-user behavior based on the IP data presented in the three samples mentioned above.

The research can be performed/done in a candidate's preferred tool (R, Python, Excel,...). All the results, description/explanation of the results, applied techniques and methods as well as the descriptions of the intermediate steps, ideas and proposals for further research should be provided in a feedback (as an attachment via email on gdjankovic@telesign.com).

2. Structure & Content of the Data

Remark 1. In the text below you'll find the explanation of all data fields, which exist in at least one out of two data sources. The data sources (*_source1* and *_source2*) don't have the same set of data fields/columns, but if a data point exists in both data sources, it will have the same name. If the data point exists in both data sources, it does not necessarily take all the available values in each data source (one out of two given data sources could be more comprehensive, refined and precise than the other one).

2.1. End-User Identity Data Fields.

fraud_label: can take the following values
- *fraud* denotes that the end-user committed a fraud,
- *legit* denotes that the end-user is a good user.

The **fraud_label** is the dependent variable, i.e. this is the real world phenomenon that we want to describe with the various IP data.

utc_time_stamp: *UTC time and date* when the transaction happened, in the form of YYYY-MM-DDT hh:mm:ss.000 (i.e. complete date plus hours, minutes, seconds and a decimal fraction of a second), for each transaction.

phone_number: *original phone number* that end-user entered in order to complete the transaction in a hashed form.

country_iso2: *Country code* of a country of origin of a **phone_number**, based on ISO-3166.

ip_address: original *end-user IP address* from which transaction was triggered in a hashed form.

start_ip: *Start IP address* of the IP range to which particular IP address belongs.

end_ip: *End IP address* of the IP range to which particular IP address belongs.

2.2. Geographic Data Fields.

Geographic data fields provide the specific details of the geo-location of the IP address:

continent: The *continent* in which the IP address is located.
continent $\in$ { Africa, Antarctica, Asia, Australia, Europe, NorthAmerica, Oceania (Melanesia, Micronesia, Polynesia), SouthAmerica }.

country: The *full country name*, where the coverage is provided for every country in the ISO-3166 alpha-2 code system.

country_code: The *ISO2 country code*, indicating the name of the origin country of an IP address, based on ISO-3166.

country_cl: *country confidence level* represents a likelihood that the user is in the location identified in the **country** field.
country_cl $\in [0, 1]$.

region: Represents *region name* information, which could contain state, city or specific territory information (e.g. Northern Territory in Australia).

state: Information for *states* and *provinces* is provided in all countries where they exist.

state_cl: *State confidence level* represents a likelihood that the user is in the location identified in the **state** field.
state_cl $\in [0, 1]$.

city: Recognizes over 150,000 distinct international *city locations*.

city_cl: *City confidence level* represents a likelihood that the user is in the location identified in the **city** field.
city_cl $\in [0, 1]$.

time_zone: Time zone is provided
- as a +/- offset from Greenwich Mean Time (GMT), so that you can calculate what time it is in the location provided and **time_zone** can take values between -11 and 13 (the time zone is derived from the **city** field if known, or from the **country** field if city is unknown), or
- in the location-format as specified by the IANA Time Zone Database (e.g. "Europe/Belgrade").

home: Indicates whether the connection is made from a residence:
TRUE: if the connection is made from a residential or private location, where mobile connections are considered private (i.e. have a TRUE value),
FALSE: if the connection is made from a place of business, organization or public place.

2.3. **Network Characteristics Data Fields.** *Network characteristics data fields* provide the characteristics of the network connection and ownership of the IP address:

connection_type: Users can connect to the Internet in several different ways:
ocx: fiber optic connections (used primarily by large backbone carriers);
tx: leased line, that is, T1, T2, T3, or T4, (used by many small/medium-sized companies); in the samples *IP_Sample_2.xlsx* and *IP_Sample_3.xlsx* this value is equal to **corporate** and **company**, respectively.
consumer satellite: consumer $\leftrightarrow$ geosynchronous or low-earth orbiting satellite links;
framerelay: frame relay circuits;
dsl: digital subscriber line broadband circuits;
cable: cable modem broadband circuits (offered by cable TV companies);
isdn: integrated services digital network technology;
dialup: consumer dial-up modem technology;
fixed wireless: fixed wireless connections;
mobile wireless: cellular network providers.
If the **connection_type** is missing, the exact connection type is unknown, and the estimated connection speed is low.

connection_speed: The *speed of the connection to the Internet* can be:
high: connection_type $\in$ {ocx, tx, framerelay};
medium: connection_type $\in$ {consumer satellite, dsl, cable, fixed wireless, isdn};
low: connection_type $\in$ {dialup, mobile wireless}.
If the **connection_speed** is missing, the exact connection type is unknown, and the estimated connection speed is low (i.e. the same as for the missing **connection_type** field).

ip_routing_type: The *IP Routing Type* (IPRT) specifies how the connection is routed through the Internet and can be used to determine how close the user is to the public IP address:
fixed: The user is connecting through a fixed-line connection. The user is likely to be at or near the location assigned to the IP.
aol/aolpop/aoldialup/aolproxy: The user is part of the AOL network. The user country in most cases can be identified, but establishing the user location below country is not possible.
pop: The user is dialing into a regional ISP and is likely to be near the IP location. Note, however, that the user might be dialing across geographical boundaries.

superpop: The user is dialing into a multi-state or multi-national ISP and is not likely to be near the IP location. Furthermore, the user might be dialing across geographical boundaries.

satellite: The user is connecting to the Internet through a consumer satellite or a backbone satellite provider, where no information about the terrestrial connection is available. In both cases, the user can be anywhere within the beam pattern of the satellite, which can span a continent or more. By definition, the satellite IPRT does not by itself indicate that the end user is connected via satellite, rather that the users traffic was routed through a satellite connection.

cache proxy: The user is using a proxy connection, either through an Internet accelerator or a content distribution service. It is possible the user is located in a different country from the IP location.

international proxy: The user is connecting through a proxy (not an anonymizer) that routes traffic from proxy multiple countries. It is possible the user is located in a different country from the IP location. In many cases, these are corporate networks that route traffic from international offices through a central point, often the corporate headquarters.

regional proxy: The user is connecting through a proxy (not an anonymizer) that routes traffic from multiple states within a single country. It is possible the user is located in a different state from the IP location. In many cases, these are corporate networks that route traffic from regional offices through a central point, often the corporate headquarters.

mobile gateway: The user is using a gateway to connect mobile devices to the public Internet. Many mobile operators, especially in Europe, serve more than one country. Therefore, it is possible the user is located in a different country from the IP location.

asn: The *Autonomous System Number* (ASN) is a globally unique number assigned to a group of networks administered by a single entity such as a Network Service Provider (NSP) or very large organization. ASNs are used to manage data routing via the Border Gateway Protocol (BGP). There are over 27,000 active ASNs. Using the ASN provides more consistency than using the carrier information, because ASNs remain static, while the specific names and ownerships of networks change.

carrier: Provides the *name of the organization that owns the ASN*. The carrier is responsible for the traffic carried on the network or set of networks designated as an Autonomous System (AS) and identified by the ASN. This field provides a more natural representation than the information provided in the **asn** field. While there are more than 27,000 active ASNs, there are fewer carriers, because a single carrier often manages several ASNs.

isp: *Internet service provider* (ISP) represents a network provider managing the network routing policy within the network range.

organization: The *Registering Organization* is the entity responsible for the actions and content associated with a given block of IP addresses. This is in contrast to the carrier, which is responsible for the routing of traffic for network blocks. Registering Organizations include many types of entities, including corporate, government, or educational entities, and ISPs managing the allocation and use of network blocks.

organization_type: Classification of the *type of* ***organization*** of the place of connection. This field provides heterogeneous categories that are useful for gaining more information about the user. Over 40 different categories for industries are provided, e.g. Insurance, Banking, Lodging, Dining, Government, Educations, Military, Recreation, Advertising, Telecommunication, as well as Internet Service Providers.

domain: Represents the *domain name* assigned to the Internet network.

sld: The *second-level domain* is the part of the domain name that precedes the top-level domain. For example, in *www.telesign.com*, telesign is the second-level domain (SLD).

tld: The *top-level domain* identifies the most general part of the domain name in a Web address. Common top-level domains (TLD) include *com*, *net*, *edu*, as well as country codes (ccTLD) like *us* (USA) and *rs* (Serbia).

hosting_facility: The *hosting facility* field contains the info whether the connection originated at a facility that provides storage, computing or telecommunication services:
TRUE: IP address is associated with a hosting facility (colocation, cloud computing, dedicated hosting, virtual private servers and web hosting),
FALSE: otherwise.

2.4. **Network Proxy Data Fields.** The *network proxy data fields*, presented in this section, provide a complete info for the identification of Internet proxies and anonymized connections, which could be used for assessment of risks and certainty when determining the location of a user.

anonymizer_status: A status is assigned to IP addresses that have been detected as a proxy. The status is an indicator, at the highest level that an IP address may be associated with an anonymizing proxy. It is a relative indicator of how recent the proxy was found to be active and the proxys category:
active: IP address was associated with an anonymizing proxy within the last month.
suspect: IP address was associated with an anonymizing proxy within the last 3 months, but not the last month.
inactive: IP address was associated with an anonymizing proxy within the last 6 months, but not in the last 3 months.
private: IP addresses labeled as *private* contain anonymous proxies that are not publicly accessible. These addresses usually belong to commercial ventures that sell anonymity services to the public (e.g. Hotspot Shield, CyberGhost).
If the anonymizer_status is vacant, then there is no specific evidence that the IP address has been associated with an anonymous proxy. After six months of inactivity an IP address ages off the **anonymizer status** list.

proxy_last_detected: Provides the most recent date on which is confirmed the proxy was active or served as a private proxy. The format of the **proxy_last_detected** date is mm-dd-yyyy. **proxy_last_detected** provides additional info for **anonymizer_status** filed and can be used to detect high-risky behavior of the user.

proxy_type: is the network or protocol utilized by the server to proxy the user connection. **Proxy_type** can be classified as:
http: The proxy uses the HTTP protocol and has open ports which are accessible by any Internet user.
service: The proxy is operated by an organization (often for profit) that provides access to subscribers as a service. The proxy is one of an array of proxies (often internationally distributed) that are part of a Virtual Private Network (VPN) that subscribers connect to by installing an application. The network may have different proxy locations or bandwidth options depending on the users membership level (paid or free).
socks: The proxy uses the SOCKet Secure (SOCKS) protocol and has open ports which are accessible by any Internet user.
tor: The proxy is part of the onion router (Tor) network. Encrypted user Internet traffic is routed through a regularly changing series of nodes operated by volunteers.

web: The proxy operates through the use of an Internet web browser. Users navigate to the web proxy website, enter the URL of the site they actually wish to visit, and the contents of the requested URL are returned by the web proxy website within the browser.

vpn: Anonymizing VPN services, which offer users a publicly accessible VPN for the purpose of hiding their IP address.

pub: Public proxies are services, which make connection requests on a user's behalf. These differ from VPNs in that the proxies usually have limited functions compare to VPNs.

dch: Hosting provider, data center or content delivery network can serve to provide anonymity, but the level of anonimity is low.

If the proxys type could not be identified the value **proxy_type** is missing.

proxy_level: The level describes the degree of concealment provided by the use of the proxy. While all proxies act as an intermediary between the user and requested website, proxies provide differing levels of obfuscation of the users originating IP address. Levels of obfuscation include:

anonymous: The proxy masks the end users IP address, but does not conceal that it is a proxy.

distorting: The proxy masks the end users IP address and does not conceal that it is a proxy. However, the end users IP address is replaced with a random IP address, thus there is a degree of subterfuge.

elite: The proxy masks the end users IP address and conceals that it is a proxy. The proxy appears to be an actual end user.

transparent: The proxy does not mask the end users IP address, nor does it conceal that it is a proxy. These proxies are typically used for information cashing and to provide joint access to Internet for multiple computers.

proxy_is_legitimate: This field will take value

- 1, if the network is a legimate proxy,
- 0, otherwise.

proxy_is_anonymous: This field will take value

- 1, if the IP address belongs to any sort of anonymous network,
- 0, otherwise.

proxy_is_anonymous_vpn: This field will take value

- 1, if the IP address belongs to an anonymous VPN system,
- 0, otherwise.

proxy_is_hosting_provider: This field will take value

- 1, if the IP address belongs to a hosting provider,
- 0, otherwise.

proxy_is_public: This field will take value

- 1, if the IP address belongs to a public proxy,
- 0, otherwise.

proxy_is_tor: This field will take value

- 1, if the IP address is a Tor exit node,
- 0, otherwise.

E-mail address: gdjankovic@telesign.com