

San Francisco Crime Classification

Kaggle competition

Łukasz Rados, Wojciech Kusa

Wydział Fizyki i Informatyki Stosowanej
Akademia Górnictwo-Hutnicza w Krakowie

26 stycznia 2016

1 Wprowadzenie

Celem projektu było stworzenie oprogramowania pozwalającego dokonać klasyfikacji zgłoszeń na podstawie danych czasoprzestrzennych z raportów policyjnych dla miasta San Francisco w Stanach Zjednoczonych. Pełen opis projektu, wraz z danymi wejściowymi znajduje się na portalu kaggle, pod adresem: www.kaggle.com/c/sf-crime/.

2 Dane

Zbiór danych zawiera incydenty zgłoszone policji w San Francisco pomiędzy 01.01.2003r. a 13.05.2015r.. Podzielony jest na dwie podgrupy (prawie równoliczne, w każdej po około 850 tysięcy elementów) :

- zbiór treningowy – zawierający zgłoszenia z tygodni parzystych,
- zbiór testowy – zawierający zgłoszenia z tygodni nieparzystych.

Przykładowe wiersze danych treningowych znajdują się na Rysunku 1. Dane składają się z następujących pól:

- Dates – znacznik czasu przestępstwa
- DayOfWeek – dzień tygodnia
- PdDistrict – nazwa departamentu policji odbierającego zgłoszenie
- Address – przybliżony adres przestępstwa
- X – długość geograficzna
- Y – szerokość geograficzna

- Category – kategoria przestępstwa (tylko dla zbioru treingowego). Jest to zmienna, którą należało przewidzieć w wyniku działania algorytmu
- Descript – szczegółowy opis przestępstwa (tylko dla zbioru treingowego)
- Resolution – jaki był wynik działania policji (tylko dla zbioru treingowego)

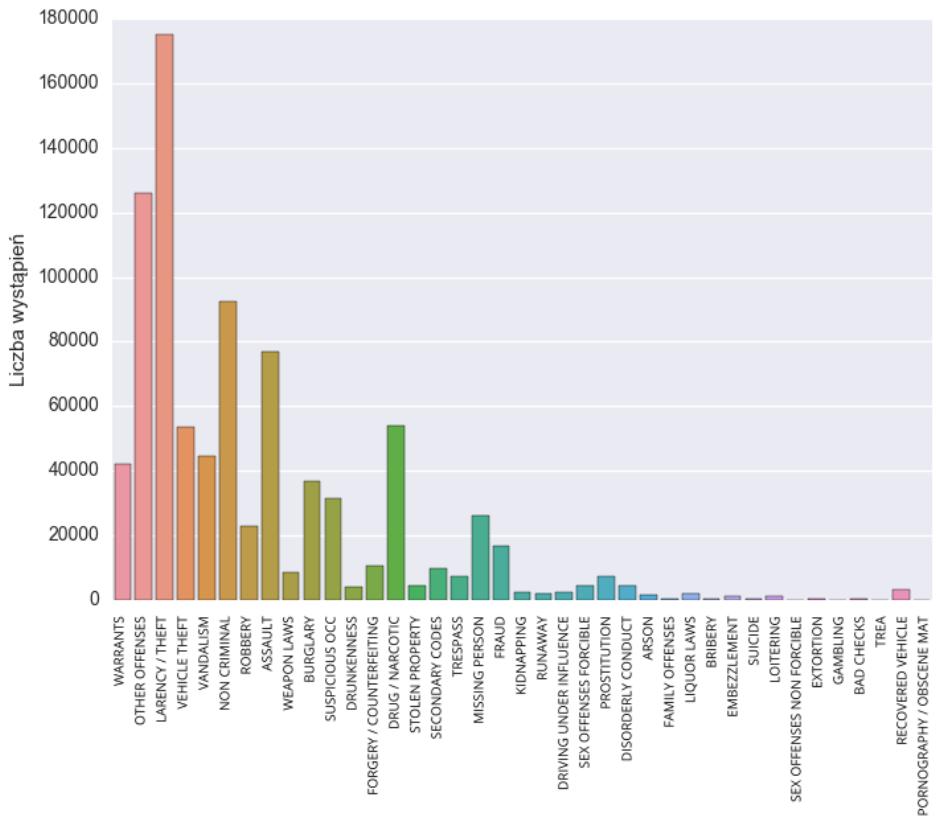
2003-01-07 07:52:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	5TH ST / SHIPLEY ST	-122.402843	37.779829
2003-01-07 04:49:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Tuesday	TENDERLOIN	ARREST, BOOKED	CYRIL MAGNIN STORTH ST / EDDY ST	-122.408495	37.784452
2003-01-07 03:52:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	OFARRELL ST / LARKIN ST	-122.417904	37.785167
2003-01-07 03:34:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	DIVISADERO ST / LOMBARD ST	-122.442650	37.798999
2003-01-07 01:22:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	900 Block of MARKET ST	-122.409537	37.782691
2003-01-06 23:30:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	BAYVIEW	ARREST, BOOKED	REVERE AV / INGALLS ST	-122.384557	37.728487
2003-01-06 23:14:00	WARRANTS	WARRANT ARREST	Monday	CENTRAL	ARREST, BOOKED	BUSH ST / HYDE ST	-122.417019	37.789110
2003-01-06 22:45:00	WARRANTS	WARRANT ARREST	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:45:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:19:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	GEARY ST / POLK ST	-122.419740	37.785893
2003-01-06 21:54:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	SUTTER ST / POLK ST	-122.420120	37.787757

Rysunek 1: Przykładowe dane treningowe. Źródło: <https://www.kaggle.com/c/sf-crime/data>

2.1 Wstępna analiza zbioru treningowego

Zbiór treningowy zawiera ponad 878 tys. zgłoszeń z okresu od 1 stycznia 2003r. do 13 maja 2015r. Przestępstwa pogrupowane zostały na 39 kategorii, których liczności przedstawiono na Rysunku 2. Większość zdarzeń zaliczona została do kategorii *Larceny / Theft* (ang. kradzież), *Assault* (ang. napaść) i *Drug / Narcotic* (ang. narkotyki) oraz grup gromadzących pozostałe zdarzenia (*Non-Criminal* oraz *Other Offences*).

Podczas analizy zbioru danych znaleziono silną zależność pomiędzy godziną a liczbą przestępstw: najmniej zgłoszeń odnotowano w godzinach 3:00 – 7:00, zaś najniebezpieczniejsze są godziny 15:00 – 20:00 (Rysunek 3). Ponadto zauważono, że rozkład zgłoszeń w czasie różni się w zależności od wybranej kategorii. Przykładem mogą być rozkłady godzinne zgłoszeń związanych z prowadzeniem po spożyciu alkoholu oraz przestępstwa narkotykowe: te pierwsze zdarzają się głównie pomiędzy godziną 20:00 a 3:00, zaś te drugie pomiędzy 12:00 a 20:00. W związku z tym prawdopodobieństwo, że przestępstwo popełnione około godziny 16 jest związane z alkoholem jest dużo mniejsze, niż szansa



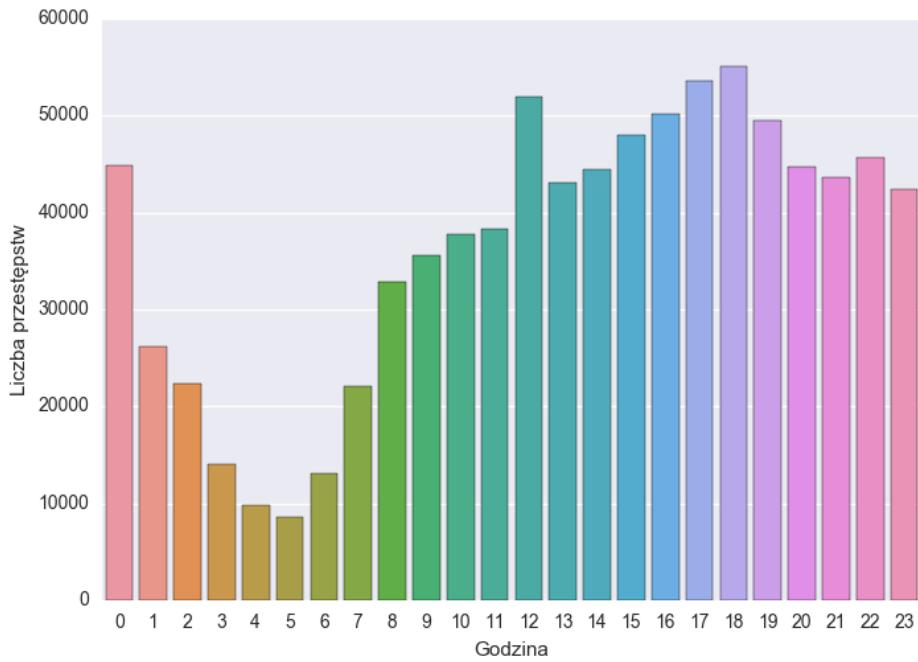
Rysunek 2: Zgłoszenia podzielone według przypisanej kategorii.

jego związku z narkotykami. Na Rysunkach 4a – 4f przedstawiono sześć przykładowych rozkładów.

Na podstawie kolumny *Address* (zawierającej przybliżony adres zgłoszenia) ustalono, czy wykroczenie miało miejsce na skrzyżowaniu dwóch ulic. Okazuje się, że niektóre z przestępstw (podejrzane zachowanie, prostytucja, jazda pod wpływem alkoholu czy sprzedaż narkotyków) są zazwyczaj powiązane ze skrzyżowaniami (w przypadku prostytucji około 80% zanotowanych przypadków miało miejsce na skrzyżowaniach). Wysoki odsetek zatrzymań za jazdę pod wpływem alkoholu na rogu ulic związany jest zapewne z rozmieszczeniem patroli policyjnych. Dane dla wszystkich przestępstw zamieszczono na wykresie 5.

Na rysunku 6b znajduje się mapa San Francisco z zaznaczonymi wszystkimi przestępstwami podzielonymi ze względu na posterunek odbierający zgłoszenie.

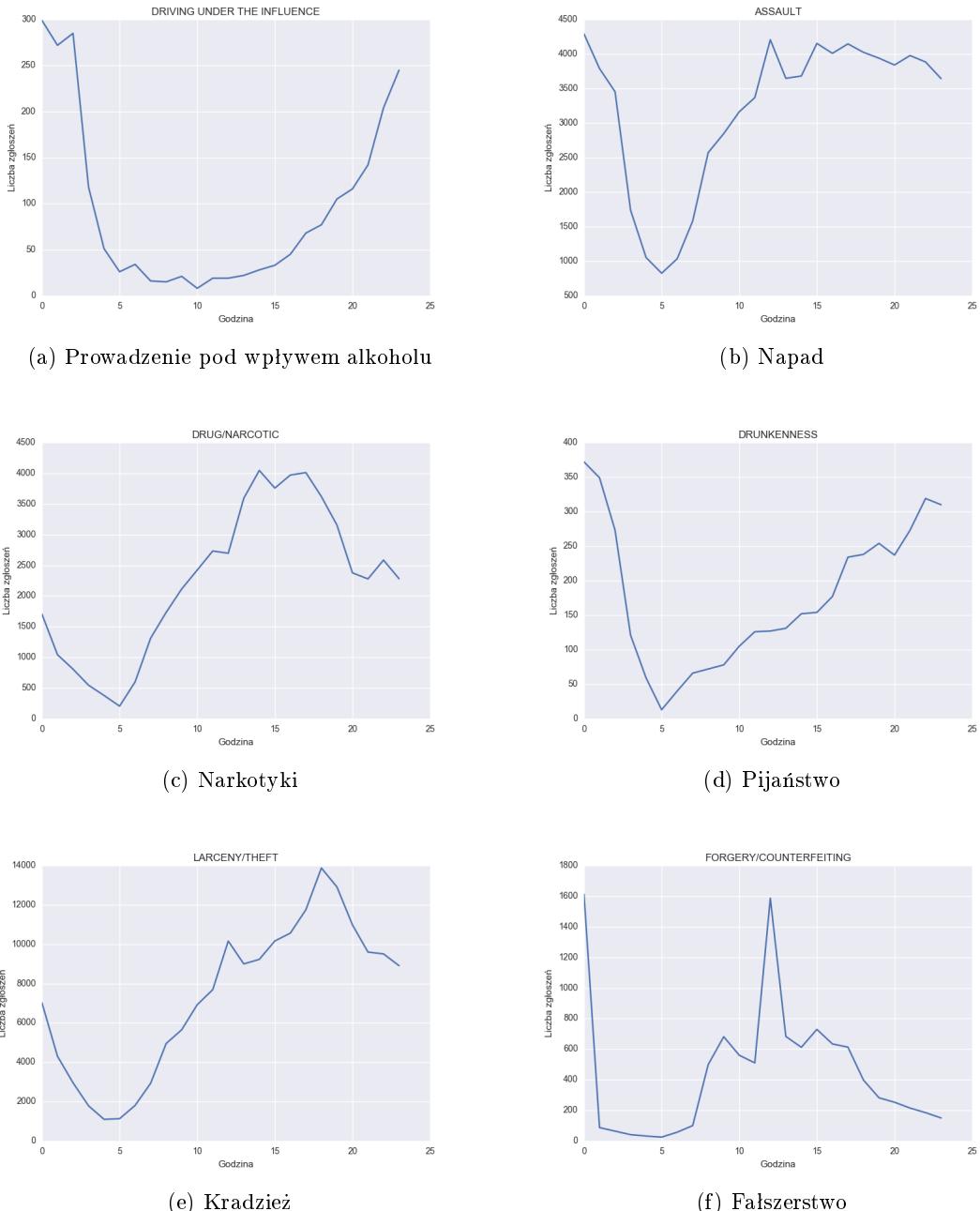
W trakcie analizy zbioru uczącego i przygotowywania wykresów, zauważono kilka cie-



Rysunek 3: Zgłoszenia podzielone według godziny w ciągu dnia.

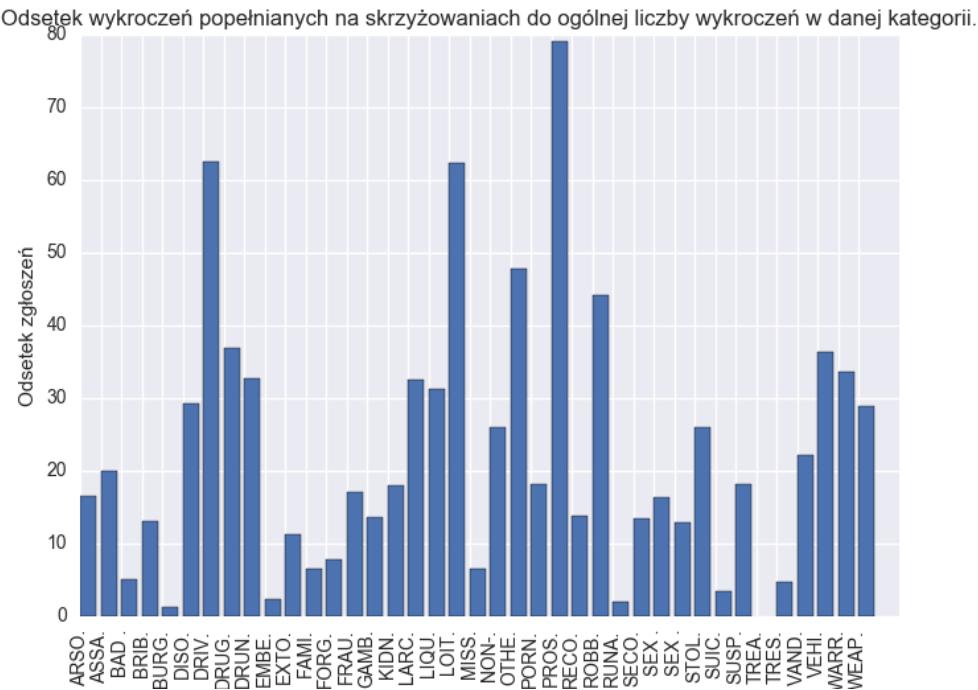
kawych (chociaż niezwiązanych z tematem pracy) zależności. Niektóre z nich przytoczono poniżej:

- Około 80% przypadków prostytucji miało miejsce na skrzyżowaniach (rysunek 5), co więcej – większość przypadków prostytucji odnotowano w dwóch skupiskach (rysunek 6c);
- Włamania są najczęstsze w godzinach 9 - 18 (wtedy, gdy większość osób jest w pracy);
- Zatrzymania związane ze spożyciem alkoholu zaczynają się pojawiać od godziny 6 rano i rosną systematycznie aż do północy;
- Największa część przypadków agresji domowej notowana jest w godzinach 13 - 15, co może być związane z końcem tzw. pierwszej zmiany;
- Kradzieże kieszonkowe mają miejsce głównie w okolicach godzin szczytu (16 - 18);
- Samochody są najczęściej kradzione między 18 a 22, zaś odzyskiwane – w godzinach południowych;
- Na 53781 zgłoszonych kradzieży samochodu przypada tylko 3138 odzyskanych pojazdów (około 6%).



Rysunek 4: Przykładowe zależności pomiędzy liczbą zgłoszeń a godziną.

Zauważone zależności pozwalają przypuszczać, że między poszczególnymi atrybutami kolejnych rekordów, zidentyfikować można subtelne powiązania.



Rysunek 5: Odsetek wykroczeń popełnianych na skrzyżowaniach do ogólnej liczby wykroczeń w danej kategorii.

2.2 Zastosowane deskryptory

Wstępna analiza zbioru treningowego pozwoliła wykluczyć z dalszej pracy następujące kolumny: *Address* (z którego wyciągnięto tylko informację o skrzyżowaniu), *Dates* (podzielono na rok, miesiąc, godzinę i minutę dnia), *Resolution* oraz *Descript* (kolumny zawierają w żaden sposób nieusystematyzowany tekst).

Dane poddane zostały preprocessingowi celem poprawy brakujących rekordów, a następnie, poza podstawowymi zmiennymi przedstawionymi w sekcji 2, przygotowano dodatkowe deskryptory mające pomóc wytrenować model. Poniżej zostały opisane najważniejsze z nich, mające istotny wpływ na poprawę działania modelu:

- $Dates.Hours \cdot 60 + Dates.Minutes$ - Liczba z zakresem 0, 1440 opisująca, w której minucie dnia zostało dokonane zgłoszenie
- $X \cdot Y$ - wskazuje na nieliniową korelację długości i szerokości geograficznej
- $X + Y$ - jak wyżej, zmienna wskazująca na korelację długości i szerokości geograficznej

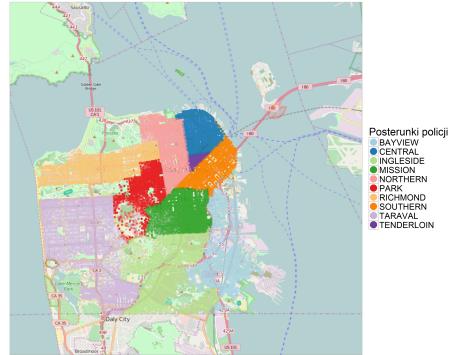
- informacja czy przestępstwo zostało dokonane na skrzyżowaniu / rogu ulicy - wyciągnięta ze zmiennej Address
- informacja czy przestępstwo zostało dokonane w bloku - wyciągnięta ze zmiennej Address
- $DayOfWeek + Dates.Hour$ - powiązuje godzinę zdarzenia z dniem tygodnia
- $DayOfWeek \cdot Dates.Hour$ - jak wyżej, powiązuje godzinę zdarzenia z dniem tygodnia

Mapa dystryktów podzielona ze względu na przestępstwa dla top 8 przestępstw.



(a) Mapa San Francisco z naniesionymi 8 najczęściej występującymi rodzajami zgłoszeń. W przypadku wyświetlania wszystkich rodzajów przestępstw, otrzymano w praktyce różnokolorową mozaikę, bardzo trudną do interpretacji.

Mapa dystryktów podzielona ze względu na posterunki policji

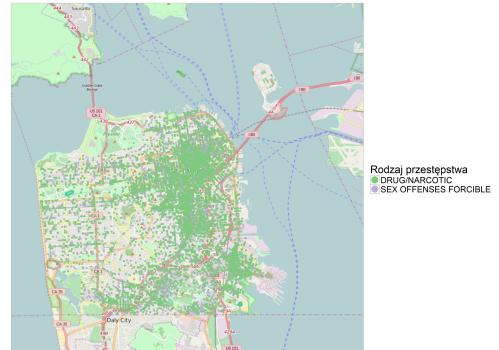


(b) Przestępstwa podzielone ze względu na posterunek odbierający zgłoszenie.



(c) Wyraźnie widoczna kumulacja przypadków prostytucji w dwóch miejscach w mieście (pozostałe kategorie rozkładają się mniej więcej równomiernie na całej mapie).

Mapa dystryktów podzielona ze względu na przestępstwa



(d) Jeden z nielicznych przypadków korelacji widocznej na podstawie mapy: powiązanie przestępstw narkotykowych z wymuszeniami seksualnymi.

Rysunek 6: Dane treningowe nанiesione na mapę San Francisco.

3 Zastosowane algorytmy

3.1 Lasy losowe - Random Forest

Algorytm Lasów Losowych (ang. *Random Forest Algorithm, RF*) to metoda klasyfikacji oparta na koncepcji bagging (podział zadania na podzbiory złożone z losowo wybranych rekordów zbioru treningowego, każdy podzbiór analizowany jest przez niezależnego eksperta). Wyniki otrzymuje się w wyniki głosowania ekspertów). Spośród całego zbioru treningowego D wybieranych jest N podzbiorów, każdy o rozmiarze m (gdzie m jest mniejsze od rozmiaru zbioru treningowego). Dla każdego z podzbiorów tworzone jest drzewo decyzyjne (algorytm korzysta więc z N drzew decyzyjnych). Ostateczna wynik klasyfikacji otrzymywany jest poprzez wybranie klasy zwróconej przez największą liczbę drzew.

Do zalet algorytmu zaliczyć można jego skalowalność (w miarę dostępności pamięci operacyjnej), małe prawdopodobieństwo przeuczenia oraz prostotę implementacji. W programie wykorzystano implementację algorytmu dostępną w bibliotece `sklearn`.

3.2 Generalized Linear Model

Uogólniony model liniowy (ang. *Generalized Linear Model, GLM*) jest rozszerzeniem zwykłej regresji liniowej pozwalającej na modelowanie zmiennych posiadających inny niż normalny rozkład błędu. GLM zezwala aby model liniowy był skorelowany z odpowiedzią poprzez funkcję łączającą (ang. *link function*) oraz wielkość wariancji dla każdego pomiaru być funkcją jego przewidywanej wartości.

Do klasyfikacji przestępstw wybrano dwumianowy rozkład błędu liczący ilość wystąpień danego przestępstwa na zasadzie binarnej tak, nie. Jako link function zastosowano funkcję logitową.

4 Implementacja

Kody źródłowe zaimplementowanych modeli znajdują się w repozytorium on-line pod adresem www.github.com/WojciechKusa/sf_crime.

4.1 Lasy losowe - Random Forest

Model wykorzystujący algorytm Lasów Losowych zaimplementowany został w języku programowania Python z wykorzystaniem bibliotek `pandas`, `sklearn` oraz `seaborn` (do wizualizacji). Ze względu na ograniczone zasoby sprzętowe (program uruchamiany był na komputerze z 8Gb pamięci RAM), liczność lasu została ustalona na poziomie 52 drzew. Większa liczba drzew zapewne pozwoliłaby na osiągnięcie lepszego wyniku (choć wą-

pliwe, by była to poprawa rzędu większego niż 0.1). W modelu uwzględniono następujące zmienne:

- $PdDistrict$
- X
- Y
- $DayOfWeek$
- $Dates.Year$
- $Dates.Month$
- $Dates.Hours$
- $Dates.Hours \cdot 60 + Dates.Minutes$
- $Block$ – informacja (0/1) czy wydarzenie miało miejsce w bloku
- $Corner$ – informacja (0/1) czy wydarzenie miało miejsce na skrzyżowaniu

4.2 Generalized Linear Model

Model wykorzystujący GLM zaimplementowany został w języku R z wykorzystaniem bibliotek MASS, readr, rpart oraz caret. Do ostatecznego modelu wzięte zostały następujące zmienne:

- $PdDistrict$
- X
- Y
- $X \cdot Y$
- $X + Y$
- $DayOfWeek$
- $Dates.Year$
- $Dates.Month$
- $Dates.Hours$
- $Dates.Hours \cdot 60 + Dates.Minutes$
- $DayOfWeek + Dates.Hour$
- $DayOfWeek \cdot Dates.Hour$
- $AddressType$ – informacja czy zdarzenie miało miejsce na skrzyżowniu czy w bloku

5 Ewaluacja oraz wyniki

5.1 Ocena modeli

Ocena modeli dokonywana była na podstawie wielo-klasowej straty logarytmicznej (ang. *multi-class logarithmic loss*). Dane zawierały prawdopodobieństwa wystąpienia danego przestępstwa przy podanych danych czasoprzestrzennych.

5.2 Uzyskane wyniki

W przypadku zastosowania modelu GLM uzyskany wynik to 2.51443 punkta co na dzień 24.01.2016r. dało 267 miejsce na około 1200 uczestników. Dla modelu RF otrzymano wynik 5.70246 co plasuje go na 889 miejscu. Wyniki te można uznać za satysfakcjonujące gdy weźmie się pod uwagę fakt, że najlepszy aktualnie wynik to 2.05079 a pierwsze 900 rezultatów to wyniki poniżej 10 punktów. Wynik dla rozwiązania losowego to 26 punktów.

6 Podsumowanie

Zadanie *San Francisco Crime Classification* nie należy do łatwych – już sama analiza mapy 6a pozwala stwierdzić, że już dla 8 najczęściej raportowanych incydentów operowanie na danych czasoprzestrzennych jest trudne. Wyniki pierwszych 100 zawodników dzielą dziesiąte części punktów, co świadczy o dużej i wyrównanej rywalizacji. Konkurs uruchomiony został 2 czerwca 2015 roku, co oznacza że zespoły biorące w nim udział miały ponad pół roku na dopracowanie swoich rozwiązań.

W tym kontekście, dla zastosowanych modeli GLM i RF uzyskano satysfakcjonujące rezultaty. Potwierdzają to wyniki w systemie konkursowym plasujące modele odpowiednio w top 300 oraz top 900 z wynikami punktowymi niewiele odstającymi od najlepszych rezultatów.

Dalszym rozwojem algorytmów mogłoby być stworzenie rozkładów gęstości prawdopodobieństwa i poszukanie korelacji pomiędzy przestępstwami co pozwoliłoby związać kilka zmiennych i (być może) uprościć model odpowiedzialny za predykcję z 39 zmiennych. Ponadto ciekawym i wartym przetestowania rozwiązaniem mogłoby być faworyzowanie częściej powtarzających się przestępstw a zerowanie prawdopodobieństwa dla tych rzadko występujących.

W przypadku rozwiązania napisanego w Pythonie, warto byłoby sprawdzić zachowanie algorytmu dla lasu złożonego z większej liczby drzew (np. 500 zamiast używanych 52).