**Analiza Danych Przestępczości w San Francisco**

**Wprowadzenie**

San Francisco jest znane z wielu rzeczy, w tym z niesławnej wyspy Alcatraz. Dzisiaj, miasto jest bardziej znane ze swojej branży technologicznej niż ze swojej przestępczej historii. Ten projekt analizuje 12 lat raportów o przestępstwach z różnych dzielnic San Francisco, aby stworzyć model przewidujący rodzaj przestępstwa na podstawie czasu i lokalizacji.

**Podjęte Kroki**

- Data Wrangling: Ocena jakości danych i wykonanie niezbędnych działań w celu oczyszczenia zestawu danych.  
- Data Exploration: Zrozumienie zmiennych i wyciągnięcie wniosków na temat danych.  
- Feature Engineering: Tworzenie dodatkowych zmiennych na podstawie istniejących.  
- Data Normalization and Transformation: Przygotowanie zestawu danych do algorytmów uczących.  
- Creation of Training and Testing Data: Ocena wydajności modeli i dostrajanie ich hiperparametrów.  
- Model Selection and Evaluation: Stworzenie modelu, który przewiduje prawdopodobieństwo wystąpienia każdego typu przestępstwa na podstawie lokalizacji i daty.

**Metryki**

Najbardziej odpowiednią metryką oceny dla tego typu problemów jest multi-class logarithmic loss.

**Analiza Danych**

Zestaw danych ma formę tabelaryczną i zawiera dane chronologiczne, geograficzne oraz tekstowe, obejmujące incydenty pochodzące z systemu raportowania incydentów kryminalnych SFPD:  
- Dates: Data (sprecyzowana do sekund)  
- Category: Kategoria incydentu przestępstwa (target variable)  
- Descript: Szczegółowy opis incydentu przestępstwa  
- DayOfWeek: Dzień tygodnia  
- PdDistrict: Nazwa dzielnicy departamentu policji  
- Resolution: Rozwiązanie  
- Address: Przybliżony adres  
- X: Długość geograficzna (Longitude)  
- Y: Szerokość geograficzna (Latitude)

**Trenowanie**

|  |  |  |
| --- | --- | --- |
| Algorithm | Parameters | Logloss |
| Stochastic Gradient Descent | Default Parameters | 20.10143 |
| K-Nearest Neighbors | Default Parameters | 26.25314 |
| HistGradientBoostingClassifier | Default Parameters | 6.59939 |
| XGBoost | Default Parameters | 2.29494 |
| LightGBM | Default Parameters | 2.59236 |
| Random Forest | Default Parameters | 5.03352 |

Ostatecznie, spośród algorytmów, które osiągnęły wynik poniżej 3.0, zdecydowaliśmy się pracować z XGBoost i LightGBM ze względu na ich efektywność i wszechstronność w dostrajaniu hiperparametrów.

**Tunowanie Hiperparametrów**

Próbowaliśmy dostrajać hiperparametry przy użyciu Random Grid Search, ale postanowiliśmy skorzystać z optymalizacji bayesowskiej. Optymalizacja bayesowska to metoda optymalizacji hiperparametrów, która buduje probabilistyczny model funkcji celu i wykorzystuje go do wyboru najbardziej obiecujących hiperparametrów do przetestowania. Dzięki zastosowaniu optymalizacji bayesowskiej możemy bardziej efektywnie dostrajać nasze modele XGBoost i LightGBM, co prowadzi do lepszych wyników predykcyjnych przy mniejszej liczbie prób.

**Pipeline'y do Przetwarzania Danych**

W celu skutecznego przetwarzania danych i porównania różnych podejść, zastosowaliśmy trzy różne pipeline'y:  
1. Pipeline z rozdzielonymi datami: Rozdzielenie dat na oddzielne kolumny reprezentujące dni, miesiące i godziny.  
2. Pipeline z funkcjami trygonometrycznymi: Zastosowanie funkcji trygonometrycznych (sinus i cosinus) do przekształcenia dni, miesięcy i godzin.  
3. Pipeline z funkcjami trygonometrycznymi i bucketowaniem: Połączenie funkcji trygonometrycznych z techniką grupowania (bucketing) współrzędnych geograficznych.

**Ewaluacja**

Ewaluujemy tylko model z najlepszym wynikiem 2.29785 tj. XGBoost wykorzystujący pipeline 2. Obecnie mamy ograniczenie do 10 ewaluacji ze względu na ograniczenia sprzętowe. Aby przeprowadzić więcej ewaluacji i szybciej znaleźć optymalne hiperparametry, musimy zdobyć większą moc obliczeniową poprzez lepszy sprzęt lub skorzystanie z usług chmurowych.