Team Project Proposal: Group-13

Team Member Details

1. Meixiang Du (md480): meixiang.du@duke.edu
2. Revanth Chowdary Ganga (rg361): revanthchowdary.ganga@duke.edu
3. Suim Park (sp699): suim.park@duke.edu
4. Titus Robin Arun (tra29): robin.arun@duke.edu

Introduction

In total, we conducted a search across 18 datasets from various sources, covering topics ranging on marketing, tourism, airline industry, weather conditions, education systems and sports datasets. Based on criteria such as data clarity, volume, and data condition, we hereby represent the winners as:

1. Los Angeles Crime Data
2. WNBA Players Data
3. Iranian Churn Data

Dataset 1: Los Angeles Crime Data (2020 to Present)

Introduction

Our top selection is a record of the incidents of crime in Los Angeles from `2020` to present day. This dataset was created by transcribing the records which are written on paper and is updated on a weekly basis.

Data Source

The data originates from the crime records of the Los Angeles Police Department (LAPD). The data can be downloaded as a csv or other file formats from the following Link.

Research Questions

1. What Are the Factors That Influence the Outcome of an Investigation?

We aim to convert ‘Status’ and ‘Status Desc’ columns in the crime dataset into binary values (solved and open cases) for outcome analysis. We’ll explore potential relationships between variables such as time gap between crime occurrence and reporting, location of crime, demographic data of victim and the outcome variable.

2. What Are the Indicating Factors That Influence the Number of Crimes Committed?

We aim to identify leading indicators for crimes by aggregating the variables including but not limited to location, demographic data. Using these variables we'll determine crime counts (the outcome) and try to predict the crime rates or the likelihood of crime occurrence.

Data Description

The Dataset is updated on a weekly basis and contains, at the time of downloading the latest dataset (20-Sep-2023) about `` `0.8 million` `` records.

the dataset contains the following columns:

- `DR_NO` : Division of Records Number
- `Date Rptd` : Date Crime was recorded in MM/DD/YYYY
- `DATE OCC` : Date Crime occured MM/DD/YYYY
- `TIME OCC` : Time of crime occurence in In 24 hour military time.
- `AREA` : Geographic Area code for the LAPD police disctrict
- `AREA NAME` : Name of the LAPD police district
- `Rpt Dist No` : A four-digit code that represents a sub-area within a Geographic
- `Part 1-2` : Indicates the part of the crime report
- `Crm Cd` : Indicates the crime committed. (Same as Crime Code 1)
- `Crm Cd Desc` : Defines the Crime Code provided.
- `Mocodes` : Modus Operandi: Activities associated with the suspect in commission of the crime
- `Vict Age` : Age of Vistim
- `Vict Sex` : Sex of Victim
- `Vict Descent` : Code to represent the descent of Victim
- `Premis cd` : The type of structure, vehicle, or location where the crime took place.
- `Weapon Used Cd` : The type of weapon used in the crime.
- `Weapon Desc` : Defines the Weapon Used Code provided.
- `Status` : Status of the case. (IC is the default)
- `Status Desc` : Defines the Status Code provided.
- `Crm Cd 1` : Indicates the crime committed.

- `Crm Cd 2`: May contain a code for an additional crime
- `Crm Cd 3`: May contain a code for an additional crime
- `Crm Cd 4`: May contain a code for an additional crime
- `LOCATION`: Street address of crime incident rounded to the nearest hundred block to maintain anonymity.
- `Cross Street`: Cross Street of rounded Address
- `LAT`: Latitude
- `LON`: Longitude

Please find below the `` `glimpse` `` of the Dataset after it has been loaded into R

```
Rows: 802,956
Columns: 28
$ DR_NO          <int> 10304468, 190101086, 200110444, 191501505, 191921269, 2…
$ Date.Rptd      <chr> "01/08/2020 12:00:00 AM", "01/02/2020 12:00:00 AM", "04…
$ DATE.OCC       <chr> "01/08/2020 12:00:00 AM", "01/01/2020 12:00:00 AM", "02…
$ TIME.OCC       <int> 2230, 330, 1200, 1730, 415, 30, 1315, 40, 200, 1925, 22…
$ AREA           <int> 3, 1, 1, 15, 19, 1, 1, 1, 1, 17, 1, 1, 1, 1, 1, 19, 11,…
$ AREA.NAME      <chr> "Southwest", "Central", "Central", "N Hollywood", "Miss…
$ Rpt.Dist.No    <int> 377, 163, 155, 1543, 1998, 163, 161, 155, 101, 1708, 19…
$ Part.1.2       <int> 2, 2, 2, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 2…
$ Crm.Cd         <int> 624, 624, 845, 745, 740, 121, 442, 946, 341, 341, 330, …
$ Crm.Cd.Desc    <chr> "BATTERY - SIMPLE ASSAULT", "BATTERY - SIMPLE ASSAULT",…
$ Mocodes        <chr> "0444 0913", "0416 1822 1414", "1501", "0329 1402", "03…
$ Vict.Age       <int> 36, 25, 0, 76, 31, 25, 23, 0, 23, 0, 29, 35, 41, 0, 24,…
$ Vict.Sex       <chr> "F", "M", "X", "F", "X", "F", "M", "X", "M", "X", "M", …
$ Vict.Descent   <chr> "B", "H", "X", "W", "X", "H", "H", "X", "B", "X", "A", …
$ Premis.Cd      <int> 501, 102, 726, 502, 409, 735, 404, 726, 502, 203, 101, …
$ Premis.Desc    <chr> "SINGLE FAMILY DWELLING", "SIDEWALK", "POLICE FACILITY"…
$ Weapon.Used.Cd <int> 400, 500, NA, NA, NA, 500, NA, NA, NA, NA, 306, 511, NA…
$ Weapon.Desc    <chr> "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)", "UNKN…
$ Status         <chr> "AO", "IC", "AA", "IC", "IC", "IC", "IC", "IC", "IC", "…
$ Status.Desc    <chr> "Adult Other", "Invest Cont", "Adult Arrest", "Invest C…
$ Crm.Cd.1       <int> 624, 624, 845, 745, 740, 121, 442, 946, 341, 341, 330, …
$ Crm.Cd.2       <int> NA, NA, NA, 998, NA, 998, 998, 998, 998, NA, NA, NA, NA…
$ Crm.Cd.3       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
$ Crm.Cd.4       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
$ LOCATION       <chr> "1100 W  39TH                          PL", "700 S  HILL…
$ Cross.Street   <chr> "", "", "", "", "", "", "", "", "", "", "OLIVE", "", ""…
$ LAT            <dbl> 34.0141, 34.0459, 34.0448, 34.1685, 34.2198, 34.0452, 3…
$ LON            <dbl> -118.2978, -118.2545, -118.2474, -118.4019, -118.4468, …
```

Dataset 2: WNBA Players Dataset

Introduction

This dataset includes season-level advanced statistics for WNBA players, organized by team, spanning the 1997-2019 seasons. d it also includes a Composite Rating obtained from a third-party source.

Data Source

The dataset originates from Basketball-Reference.com, for this project the dataset was downloaded from the following Link

Research Questions

1. What is the primary indicator of a player's probability of winning and their total share of winnings?

We aim to identify the key factors influencing a player's total win shares (outcome variable) by examining variables such as the player's age, team, their position within the team, and their efficiency level in each respective position.

2. What distinguishes athletes who achieve long-term success in their careers from their peers?

We aim to enhance our understanding of the leading indicators that set apart athletes who enjoy enduring success in their careers from their peers investigation by encompassing factors such as player age, team affiliation, playing position within the team, efficiency levels in each position, as well as any changes in position and efficiency over different career stages.

Data Description

The Dataset contains `3883` records and following columns:

- `Player` : Player name
- `year_ID` : Season
- `Age` : Age (as of Jul. 1)
- `Tm` : Team played for
- `tm_gms` : Team's scheduled games
- `Pos` : Player's position played
- `G` : Games played
- `MP` : Minutes played
- `MP_pct` : Percentage of available minutes played
- `PER` : Player Efficiency Rating

- `Status`: Binary (1: Active, 2: Non-active)
- `TS_pct`: True Shooting Percentage
- `ThrPAr`: Three-point Attempt Rate (3PA/FGA)
- `FTr`: Free Throw Rate (FTA/FGA)
- `ORB_pc`: Offensive rebound percentage
- `TRB_pct`: Total rebound percentage
- `AST_pct`: Assist percentage
- `STL_pct`: Steal percentage
- `BLK_pct`: Block percentage
- `TOV_pct`: Turnover percentage
- `USG_pct`: Usage percentage
- `OWS`: Offensive Win Shares
- `DWS`: Defensive Win Shares
- `WS`: Total Win Shares
- `WS40`: Win Shares per 40 minutes
- `Composite_Rating`: Estimated net points added per 100 possessions
- `Wins_Generated`: Wins implied by Composite Rating

please find below the `glimpse` of the Dataset after it has been loaded into R

```
Rows: 3,883
Columns: 28
$ player_ID        <chr> "montgre01w", "williel01w", "sykesbr01w", "hayesti01w…
$ Player           <chr> "Renee Montgomery", "Elizabeth Williams", "Brittney S…
$ year_ID          <int> 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019, 2019,…
$ Age              <int> 32, 26, 25, 29, 31, 28, 23, 22, 24, 25, 24, 26, 32, 2…
$ Tm               <chr> "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", "ATL", "ATL…
$ tm_gms           <int> 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 3…
$ Tm_Net_Rtg       <dbl> -9.8, -9.8, -9.8, -9.8, -9.8, -9.8, -9.8, -9.8, -9.8,…
$ Pos              <chr> "G", "C-F", "G", "G", "F", "G", "F", "G", "F", "C", "…
$ G                <int> 34, 32, 34, 29, 33, 29, 29, 31, 28, 31, 9, 4, 1, 34, …
$ MP               <int> 949, 909, 880, 817, 767, 634, 553, 477, 389, 349, 75,…
$ MP_pct           <chr> "69.5%", "66.6%", "64.5%", "59.9%", "56.2%", "46.4%",…
$ PER              <dbl> 11.1, 16.7, 11.3, 15.1, 15.2, 8.8, 11.7, 6.4, 10.1, 8…
$ TS_pct           <dbl> 0.520, 0.521, 0.445, 0.497, 0.425, 0.384, 0.469, 0.44…
$ ThrPAr           <dbl> 0.727, 0.000, 0.308, 0.313, 0.138, 0.421, 0.007, 0.59…
$ FTr              <dbl> 0.176, 0.477, 0.259, 0.340, 0.120, 0.108, 0.417, 0.25…
$ ORB_pct          <dbl> 1.0, 11.4, 3.0, 2.7, 5.7, 1.6, 11.5, 0.4, 6.2, 10.5, …
$ TRB_pct          <dbl> 4.1, 12.1, 9.1, 5.7, 16.6, 4.5, 19.3, 2.3, 10.6, 12.9…
$ AST_pct          <dbl> 18.0, 7.8, 19.6, 21.9, 14.2, 28.1, 5.4, 17.8, 6.0, 10…
```

```
$ STL_pct          <dbl> 1.7, 1.4, 1.2, 1.8, 2.8, 1.7, 1.6, 1.9, 1.9, 0.7, 1.4…
$ BLK_pct          <dbl> 0.5, 4.7, 1.5, 0.8, 3.6, 1.1, 1.7, 0.2, 2.7, 5.5, 6.4…
$ TOV_pct          <dbl> 16.8, 12.9, 14.8, 14.0, 11.6, 11.5, 21.3, 21.9, 11.8,…
$ USG_pct          <dbl> 17.5, 15.9, 23.1, 27.1, 18.9, 26.7, 17.3, 12.3, 20.3,…
$ OWS              <dbl> 0.4, 1.6, -0.8, 0.5, -0.4, -1.4, -0.3, -0.4, -0.3, -0…
$ DWS              <dbl> 0.5, 1.0, 0.8, 0.6, 1.9, 0.4, 1.0, 0.2, 0.5, 0.4, 0.2…
$ WS               <dbl> 0.9, 2.7, 0.0, 1.0, 1.5, -1.0, 0.7, -0.2, 0.2, -0.1, …
$ WS40             <dbl> 0.039, 0.117, -0.001, 0.050, 0.076, -0.063, 0.049, -0…
$ Composite_Rating <dbl> -2.4, 0.6, -3.4, -1.5, -0.8, -5.5, -2.1, -4.6, -3.1, …
$ Wins_Generated   <dbl> 1.22, 2.51, 0.70, 1.45, 1.62, -0.14, 0.81, 0.10, 0.37…
```

Dataset 3: Iranian Churn Dataset

Introduction

The third dataset we want to use is a record of customers' churn resulted from the Iranian telecom company. All of the attributes except for churn are the aggregated data of the first 9 months.

Data Source

This dataset is randomly collected from an Iranian telecom company's database over a period of 12 months.

The dataset has been downloaded from the following Link

Research Questions

1. What are the indicators that impact the churn?

We aim to identify leading indicators for churns. By aggregating data at the variable of information from telecom company, we'll determine churns (the outcome). Since many variables can potentially influence churn, we will investigate which variables have the greatest impact on churn.

2. What are the factors that impact how long a customer stays with the carrier?

we plan to calculate the duration of customer engagement as a field of interest, and explore customer complaint behavior and purchasing behavior to identify the key indicators for identifying loyal customers.

Data Description

The Dataset contains `3150` records and following columns.

- `Call Failures` : Number of call failures
- `Complains` : Binary (0: No complaint, 1: Complaint)

- `Subscription Length`: Total months of subscription
- `Charge Amount`: Ordinal attribute (0: lowest amount, 9: highest amount)
- `Seconds of Use`: Total seconds of calls
- `Frequency of use`: Total number of calls
- `Frequency of SMS`: Total number of text messages
- `Distinct Called Numbers`: Total number of distinct phone calls
- `Age Group`: Ordinal attribute (1: younger age, 5: older age)
- `Tariff Plan`: Binary (1: Pay as you go, 2: Contractual)
- `Status`: Binary (1: Active, 2: Non-active)
- `Age`: Age of the customers
- `Customer Value`: The calculated value of customers
- `Churn`: Binary (1: churn, 0: non-churn) - Class label

please find below the `` `glimpse` `` of the Dataset after it has been loaded into R

```
Rows: 3,150
Columns: 14
$ Call..Failure           <int> 8, 0, 10, 10, 3, 11, 4, 13, 7, 7, 6, 9, 25, 4,…
$ Complains               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Subscription..Length    <int> 38, 39, 37, 38, 38, 38, 38, 37, 38, 38, 38, 38…
$ Charge..Amount          <int> 0, 0, 0, 0, 0, 1, 0, 2, 0, 1, 0, 0, 3, 1, 0, 1…
$ Seconds.of.Use          <int> 4370, 318, 2453, 4198, 2393, 3775, 2360, 9115,…
$ Frequency.of.use        <int> 71, 5, 60, 66, 58, 82, 39, 121, 169, 83, 95, 5…
$ Frequency.of.SMS        <int> 5, 7, 359, 1, 2, 32, 285, 144, 0, 2, 7, 8, 54,…
$ Distinct.Called.Numbers <int> 17, 4, 24, 35, 33, 28, 18, 43, 44, 25, 12, 17,…
$ Age.Group               <int> 3, 2, 3, 1, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3…
$ Tariff.Plan             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ Status                  <int> 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1…
$ Age                     <int> 30, 25, 30, 15, 15, 30, 30, 30, 30, 30, 30, 30…
$ Customer.Value          <dbl> 197.640, 46.035, 1536.520, 240.020, 145.805, 2…
$ Churn                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```