CS 334 Machine Learning

Spotlight Presentation

November 1st

Format

- Each project slide is set to 2 mins
- The group slide before project slide is set to 1 min for Q/A and transition

Group #16

Hospital Bill Pricing

Presented By: Ruilin Chen, Ethan Krein, Danielle Linbeck

Hospital Bill Pricing

Overview and Motivation

- Hospital pricing for the same services can vary greatly between providers regardless of quality of care
  - Difficult for patients to access and interpret published data
  - Non-compliance with transparency laws leaves gaps in the data
- Goal provide patients with accurate price estimates for specific procedures

Dataset

- DoltHub Hospital Price Transparency dataset (2021):
  - 13 features from over 1400 providers across more than 1100 hospitals in the US
  - Provides prices for each procedure by hospital along with location and payment method
- Previous work: hospital compliance rates and price changes due to transparency

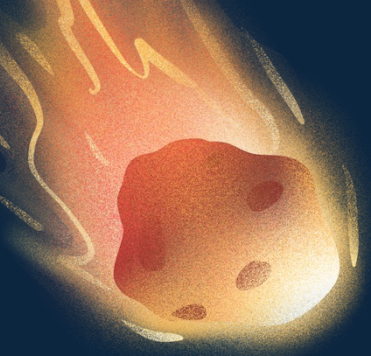
Methodology

- Preprocessing: feature trimming, standardization, one-hot encoding
- Models: K-nearest neighbors, decision trees, ensemble (bootstrap)
  - With K-fold cross validation
- Evaluation: Mean squared error

Group #15

Classification of Stars, Galaxies, and Quasars Utilizing Supervised Machine Learning via SDSS Photometric Data

Presented By: Mingxuan Liu & Yijun Liu



Classification of Stars, Galaxies, and Quasars Utilizing Supervised Machine Learning via SDSS Photometric Data

G15: Mingxuan Liu & Yijun Liu

Overview & Motivation

Problem: Quasars are difficult to classify because of indistinguishable images between stars and quasars

Two types of astrophysics data: spectroscopic (expensive, time-consuming; yet can distinct quasars from stars clearly) and photometric (cheap; unable to distinguish quasars easily)

Goal: Using photometric data to distinguish quasars

Significance & Novelty:
-Further understanding in supermassive black holes and dynamics around AGN
-Explore multi-color space

Methodology

Data: SDSS (105,783 celestial objects; 74 features--identifiers, coordinates, magnitudes in various light bands, redshift and flux measurements, signal-to-noise ratios, and separational data)

Model: DT, RF, SVM, KNN, Neural Networks, XGB

Feature Selection: domain knowledge, filter (ANOVA), wrapper (forward, backward, & RFE), embedded (LASSO/Ridge, FI scores), PCA
New Feature: multi-color space

Hyperparameter: grid/random search

Evaluation: precision, recall, f1, AUC-PR

Novelty: Only use SDSS, more models, add galaxies, add multi-color space

Group #14

Predicting Fraud in Online Payments: A Machine Learning Investigation

Presented By: Matthew Sharp, Cayra Williams, Brandon Zhang

Question + Motivation: What are Key Strategies for Mitigating Risks and Enhancing Security in the Dynamic Landscape of E-commerce Transactions?

- With a recent rise in security threats regarding mobile banking/E-commerce, our team wanted to explore the possibility of alleviating these risks through the use of machine learning models.
- We see these types of models already being implemented by companies such as Amazon, PayPal, and Stripe, however, as AI-powered security threats get more advanced, there is no doubt that these models will need to become more complex and thoroughly incorporate many of the topics we covered in CS 334.
- AI/ML based security products are an extremely rapid growing industry, with Morgan Stanley predicting a global market growth from $15 billion in 2021 and to roughly $135 billion by 2030.

Dataset: Online Payments Fraud Detection Dataset

- Features: 6M+ rows, Amount of transaction, Type of transaction, Anonymous customer and merchant IDs, Before and after balances for customer and merchant, Labels indicating fraudulent or non-fraudulent transactions
- Existing work: Work previously done on dataset, but did not include comprehensive performance evaluation nor did it compare machine learning models different models
- Source: (found here)
  - Uses simulated data created by PaySim, as described in the work of Rojas et al. (references available in proposal)

Methodology:

- Product Management System:
  - Scrum
- Data Preprocessing:
  - Train-test split for the data set
  - feature extraction
  - feature scaling using z-score standardization
- Modeling Approach:
  - knn, decision tree classifier, neutral networks
  - Comparison of efficiency, accuracy, precision, and recall using data visualization
  - Performance evaluation using metrics such as AUC-ROC and F1-score
  - Hyperparameter tuning utilizing grid search with cross-validation to counteract overfitting

Group #13

Optimized Airline Bookings

Presented By: Cashin Woo, Dhruv Naheta, Jordan Leslie

Optimized Airline Bookings

Overview & Motivation

- *Problem:* travel plans disrupted by delayed/canceled flight
- *Solution:* train model to predict delays and cancellations

Dataset: Kaggle Flight Status Prediction

- *Size:* all U.S. flights from 2018-2022; millions of samples
- *Features:* 61 columns; location, date/time, delay measurements, airline information
- *Existing work:* binary classification, DT performed best

Methodology

- *Feature selection:* mutual information or Pearson correlation
- *Preprocessing:* one-hot encoding, normalization, handle potential missing values
- *Models:* KNN, decision tree, naive Bayes, logistic regression, neural network, random forest
- *Hyperparameter tuning:* Parameter weight optimization(using grid search)
- *Evaluation:* Notable increase in prediction of delays and cancellation compared to current models

Group #12

Understanding the Risk Factors of Long COVID Using the Behavioral Risk Factor Surveillance System

Presented By: Devin Gee, Juan P. Selame Fernandez

Understanding the Risk Factors of Long COVID Using the Behavioral Risk Factor Surveillance System

Team: Devin Gee, Juan P. Selame Fernandez

Overview: The CDC's Behavioral Risk Factor Surveillance System (BRFSS) is an annual phone survey that collects information centered on health and behavior among American adults. The 2022 survey was the first year that included information relating to COVID, providing new opportunities to understand the behaviors, demographics, and health conditions linked to long COVID.

Dataset: Modeled after a Kaggle dataset that used the 2015 data

- **Source**: CDC's Population Health Surveillance Branch
- **Features**: 22 in total including but not limited to: demographic (age, sex, race, income), lifestyle (sleep amount, exercise, diet,smoking), health (BMI, chronic illnesses, cholesterol, blood pressure)
- **Samples**: 121,379 responses (pre-processing)
- **Goal**: Predict whether or not the respondent has/had long COVID and identify key predictive features across models.

Work Done So Far: Previous project have attempted to predict Long COVID largely through the use of symptom focused datasets. None have used more sociobehavioral datasets to understand the backgrounds and contexts of those with the disease.

Methodology:

- **Preprocessing**: Clean the 2022 CDC data for analysis selecting features of interest and removing invalid samples; partition into training and testing sets. Strong focus on feature selection such as filter methods (Pearson correlation) and embedded methods (Lasso and Ridge regression)
- **Model Experimentation**: Test machine learning models including KNN,Logistic Regression, Random Forest, Naive Bayes and Neural Networks.
    - K-Fold Cross Validation to improve models
- **Optimization & Evaluation:** Standardize data, and tune hyperparameters with random search. Evaluate using AUROC and AUPRC due to dataset imbalance

Group #11

Learning Online Shoppers Intention

Presented By: Bella Li, Angelina Ying, Candy Gao

Learning Online Shoppers Intention

1. Problem Description

Business Challenge

Optimize marketing decisions using user behavior data

Identify impactful factors in marketing decisions

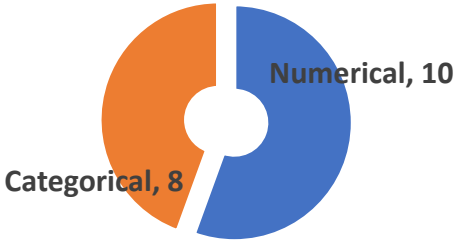
How we plan to address it

Test machine learning algorithms for predicting the influence of special day promotions on online shoppers

Expected outcome & Future impact

Form a model selection guideline for future predictions

2. Online Shoppers Purchasing Intention Dataset



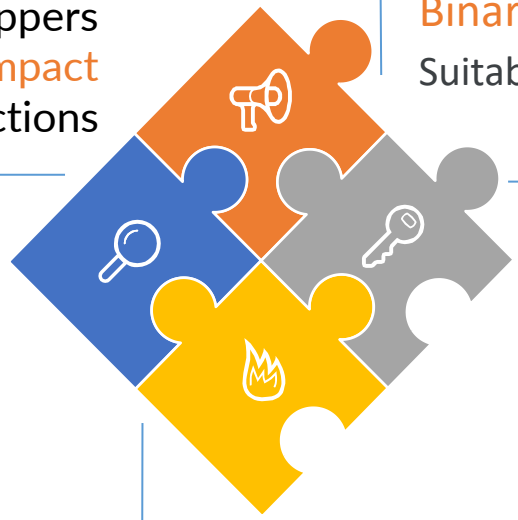
12,330 sessions

Over 1-year of time span

No missing values

Binary "Revenue" attribute indicates purchase

Suitable for comparing ML models



3. Work So Far

Previous works have shown high correlation with the features: "PageValues" and "BounceRates".

These two tend to contribute significantly to the output results. Moreover, clustering appears to have a promising result in studying this dataset.

4. Tentative Plan

Data Preprocessing

- Normalize numeric values & Convert categorical

Feature Selection

- Eliminate insignificant features
- Model Selection & Assessment

Employ various supervised learning algorithms

- Utilize various validation techniques
- Assess using quality metrics (accuracy, recall, precision, F-score, AUC, AUCPR, etc.)

Large dataset (12330 entries)
10 numerical and 8 categorical attributes

Group #10

Fingerspelling Classification

Presented By: Kevin Seo, James Song

Overview: How accurate are ML models in classifying ASL (American Sign Language) finger-spelling?

Motivation: To utilize ML to make learning ASL finger-spelling easier and more accessible for those in need.

Dataset:

We will be working with two datasets: the "*Sign Language MNIST*" dataset and the "*ASL Fingerspelling*" dataset.

- The "*Sign Language MNIST*" dataset follows the same .csv format with labels and pixel values in single rows as the well known MNIST database
  - label of the alphabet (target) with 784 pixel values (784 features) for each of the 338,313 train data and 7,172 test data.
- The "*ASL Fingerspelling*" dataset contains images of 5 subjects spelling out each alphabet more than 1000 times.

Methodology:

- Preprocessing:
  - Clean ASL Fingerspelling data from different sources into one coherent dataset
- Feature Selection:
  - Variance Threshold: pixels with low variance across all images can be removed as they might not be carrying useful information.
  - Recursive Feature Elimination with Cross-Validation (RFECV): Uses accuracy metric to recursively remove features and builds models using cross-validation.
- Models:
  - k-NN, Random Forest Classification and other neural network methods
- Hyperparameters / Evaluation:
  - Grid Search
  - Accuracy score & Confusion Matrix

Group #9

Stock Price Prediction Model

Presented By: Brian Hsu, Jason Li

Group 9 – Stock Price Prediction Model

Problem: Stock price is hard for human to predict; can ML models perform better?

Why: Recent studies found that machine learning models such as the long short-term memory (LSTM) model is good at capturing time series data such as stock price. We want to take a deeper look to use models we learned in class to predict stock price and compare their accuracies against LSTM. We will select 5 companies from different industries to train and test the model performance.

Dataset: We will get the company stock dataset from Yahoo Finance using yfinance library with a range from 2000/1/1 to 2022/12/31. Features include hourly high, low, close, volume, market cap, beta, competitors/supplier's stock price, etc. Adjusted price to consider dividends and splits. 40000 samples, 30 attributes.

Existing Work: LSTM model on stock price (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9257950)

Models: Linear Regression, Random Forest, Decision tree, Ensemble methods (bagging, boosting, stacking), Moving Average, Long short-term memory

Methodologies:

- Feature selection: Pearson correlation, Ridge regularization
- Data Preprocessing: Standard scaling, one-hot encoding for binary variables
- Train/Test Split: 0.7/0.3; K-fold validation
- Hyperparameter tunning: Grid search
- Evaluation: Mean Absolute Error and RMSE

Group #8

Kingdom Classification from DNA Codon Usage Data

Presented By: Yingrong Chen, Nathan Yoo, Andy Kim

Kingdom Classification from DNA Codon Usage Data

Question:

- Every organism uses 64 possible 3-letter DNA codons to encode 20 amino acids, but the extent to which each organism uses these codons varies based on species. Can we use codon usage data to classify organisms into different life kingdoms?

Motivation:

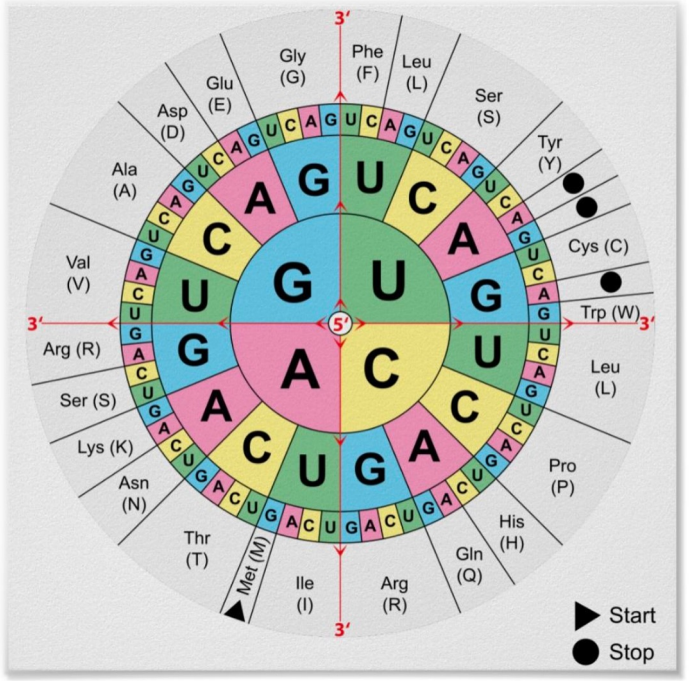
- We can gain insights into evolutionary relationships using machine learning models, replacing the slower and more subjective traditional taxonomy based on morphology and biochemistry.

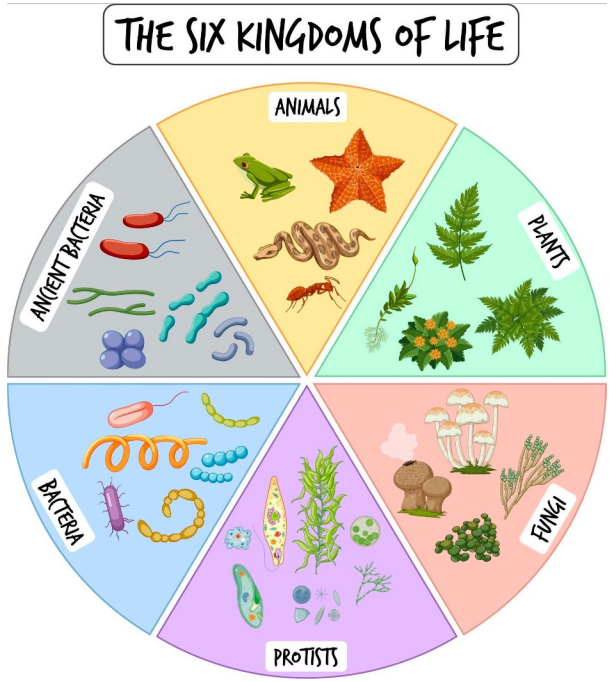
Description of the Dataset:

- Publicly available dataset on UCI Machine Learning Repository.
- 13,028 entries, 69 attributes (64 being the different codons).
- *Prediction Target*: 11 different evolutionary classes.
- *Previous work:* Ensemble “hard-voting” method of variety of models (k-NN, RF, SVM).

Methodology:

- *Preprocessing:* One-Hot encoding categorical variables. Imputing missing values. Standardizing/Min-Max scaling of numerical variables.
- *Models*: k-NN, SVM, RF, Naive Bayes, Logistic Regression, Decision Trees, Ensemble methods such as weighting voting of different models, bagging, and stacking. hyper-parameter tuning(GridSearch) through k-fold cross validation
- *Evaluation*: accuracy, precision, recall, AUC



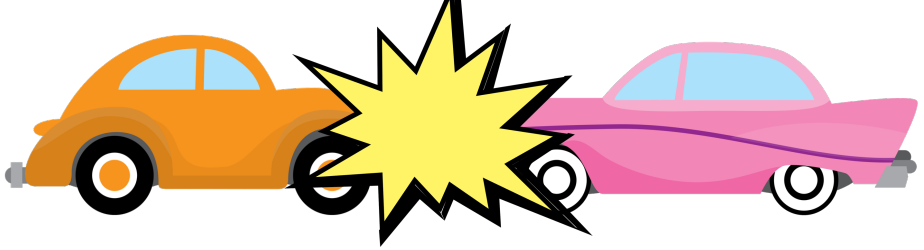


Group #7

Collision Severity Prediction

Presented By: Zoie Peng, Annie Rhoades

Collision Severity Prediction

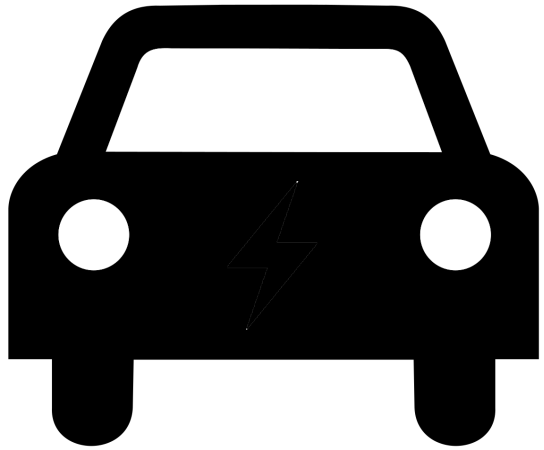


By: Zoie and Annie

- **Overview:** Using the US Traffic Accidents (2016-2023) dataset, compare the performance of different models' severity predictions
  - 10 weather features
  - 13 location features
  - 4 time features

  Over 7 million samples

  45 features per sample
- **Motivation:** Find the best model to predict accident severity based on relevant features
  - compare different models
  - use the best model to assess accident risk for certain conditions
- **Methodology:**
  - Pre-processing: feature selection, standard scaling, one-hot encoding, missing values
  - Models: K-Nearest Neighbors, Decision Tree, Naïve Bayes
  - Hyperparameter Tuning:
    - KNN: k, weights, distance metrics
    - Decision Tree: criterion, min leaf samples, tree depth
  - Cross Validation: k-fold cross validation
  - Evaluation Metric: AUPRC



Group #6

A side-by-side comparison of multiple machine learning algorithms on the detection of fraudulent credit card transaction

Presented By: Steve Li, Tianjun Zhong, Zinc Zhao

A side-by-side comparison of multiple machine learning algorithms on the detection of fraudulent credit card transaction

MOTIVATION

The effective detection of fraudulent transactions is imperative for credit card companies to ensure the security of their customers' financial assets. It is imperative to employ robust and efficient techniques to detect anomalies in credit card transactions to eliminate losses at the source.

Dataset

The dataset encapsulates credit card transactions made during September 2013 by European cardholders. It spans across two days and comprises 284,807 transactions, out of which 492 are fraudulent. For confidentiality constraints, original features were transformed and labeled V1 to V28, in addition to timestamp, transaction amount and true labels.

Methods

Multilayer Perceptron, Perceptron, Logistic Regression, k-NN, Decision Tree
We will be using k-Fold to tune hyperparameters.

Evaluation

Due to the extreme unbalance in our dataset (0.17%) fraudulent, we will be using Precision/Recall to get an accurate representation of model performance, since Accuracy will be inflated by the rare occurrence of the fraudulent class.

Group #5

Determine what attributes of an NBA player are most important to make an All-NBA team

Presented By: Louis Mullarkey

Goal: Determine what attributes of an NBA player are most important to make an All-NBA team

Motivation

- Structure of NBA contracts
- Making an All-NBA team lets you sign a large contract early
- All-NBA teams voted by 100 media members and retired players
- Ambiguity and subjectivity of All-NBA criteria
- Give players more direction in what parts of the game to focus on

Dataset:

- Seasons since 1976 from players who averaged at least 5 points per game
- BasketballRefrence.com
- Predicted class: All NBA team
- Features
  - Base metrics (ppg, apg, rpg)
  - Advanced metrics (PER, Win Shares, True Shooting%)
  - Team metrics (wins, team ppg, opponent ppg)

Methodology

- Preprocessing: Standard Scaling/Min-max Scaling
- Feature Selection: Ridge Regression/Lasso Regression
- Models: KNN, Decision Tree, Gradient Descent, Naive Bayes
- Evaluation: K folds cross validation + Accuracy/ Confusion Matrix

Group #4

Predicting Number of Owners With Steam Data

Presented By: Jonathan Kim, Microl Chen, Peter Jeong

Predicting Number of Owners With Steam Data

Jonathan Kim, Microl Chen, Peter Jeong

Overview

- Rising video game popularity over the years
- The ability to determine the success of a product is crucial
- Previous work focus and actual sales value, instead of number of games sold
- We can make further progress with new dataset and features

Dataset

- Game statistics gathered from ESD Steam by maxwell in Kaggle (2023)
- **Features**: 70000+ samples, 20 features selected
- **Predicted** Class: Categorical Classification (Estimated Owners)
- **Previous Work**: Video Games dataset gathered by Dr. Joe Cox (2017)

Methods

- **Preprocessing**
  - Label Encoding for Categorical features, Counting number of genres for the list of genres
  - Standard/Min-max scaling, handling missing values
- **Models**: K-Nearest Neighbors, Decision Tree, Neural Networks
- **Hyperparameter Tuning**: Grid Search, Ensemble Methods
- **Evaluation**: K-fold Cross Validation, Monte Carlo Cross Validation

Group #3

Identifying Players Who Hit It Out of Their Ballpark (Estimate)

Presented By: Tommy Skodje, Clay Winder

Identifying Players Who Hit It Out of Their Ballpark (Estimate)

- Overview/Motivation
  - Recent trend in baseball towards recording advanced statistics.
    - WAR (Wins Above Replacement).
    - wRC (weighted Runs Created).
  - Try to quantify the **value** of a player.
  - Even advanced stats do not record *how* an outcome happened, only *what* happened.
  - Statcast - Answers this “how”. Exit velocity, launch angle, and sprint speed.
  - **Goal - Determine which players are overvalued and undervalued based on their Statcast statistics.**
- Methodology
  - Approach: kNN, linear regression, bagging & random forest
  - Dataset: 3920 samples x 70 features = 274,400 (will use Pearson correlation for preprocessing)
  - Evaluation: k-fold cross-validation (tuning kNN hyperparameters), OOB error (tuning random forest)
  - Previous Work: Using Statcast data to find undervalued pitchers and predict wOBA (an important hitting stat)

Group #2

Point Prediction: Leveraging Historical Data and Player Comparisons for NBA Players' Points Per Game Forecasting

Presented By: William Chung, Lemar Minott

Point Prediction: Leveraging Historical Data and Player Comparisons for NBA Players' Points Per Game Forecasting

Overview and Motivation: Modern basketball has gradually increased its reliance on data-driven models in order to analyze player performances and potential. Due to our interest in the sport, we thought it would be interesting to develop a model that predicts a NBA player's Points Per Game (PPG) through the use of historical data, leveraging the unique correlations and patterns gained from comparing a player's statistics to their own previous records and those of their peers with similar characteristics.

Dataset: The Kaggle dataset we plan to use provides a collection of 12,000+ players who have been part of an NBA team's roster from 1996 to 2021. This dataset includes a wide range of feature columns, with demographic variables such as age, height, weight, place of birth, and biographical information, as well as the box score statistics such as games played, average number of points, rebounds, assists, etc.

Methodology: Models to be used: K-Nearest Neighbors (KNN) and Linear Regression (we will utilize K-fold cross validation to improve the models), Preprocessing: Standard Scaling, and Evaluation: Accuracy/Confusion Matrix.

Group #1

Heart Disease Prediction

Presented By: Jeffrey Taylor, Christopher Roebuck, Jake Brown

Fantasy Football Points Predictor

Jeffrey Taylor, Topher Roebuck, Jake Brown

Problem: Can we outperform existing predictive NFL player performance models on a week to week basis?

Motivation:

- Construct highly accurate model to predict weekly player output
- Existing predictors tend to be inaccurate/undeveloped

Preprocessing:

- Feature engineering and scaling
- Cleaning, dealing with missing values (e.g., injuries, retired players)
- LASSO and/or PCA for dimensionality reduction

Dataset:

- Weekly statistics and fantasy scoring data scraped from fantasypros.com (2018 - present)
- Over 35,000 observations and at least 30 features

Methodology:

- Optimal combination of ensemble methods (begin with LR and RF)
- Most accurately predict FPTS for the week

PLAYER	CMP	ATT	PCT	YDS	Y/A	TD	INT	SACKS	ATT	YDS	TD	FL	G	FPTS
Patrick Mahomes II (KC)	32	42	76.2	424	10.1	4	1	1	4	29	0	0	1	34.9

Thank you!