FINAL PROJECT

PREDICTIVE MODELS

By : Winnie Yang and Sophie Dang

TARGET/AIM

- Find a predictive model that estimates NYC government employee salary based on available different features
- How much a NYC employee is paid
- What influences their salary

Data Description

- "Citywide Payroll Data (Fiscal Year)" from NYC Open Data.
- Captures municipal salary expenditures and analyzes salary determinants.
- Focus Variables: Agency name, title description, pay basis, hours, and gross pay.
- Scope: Contains over 6 million records, providing a representative view of NYC's workforce compensation.

	Fiscal Year	Payroll Number	Agency Name	Last Name	First Name	Mid Init	Agency Start Date	Work Location Borough	Title Description	Leave Status as of June 30	Base Salary	Pay Basis	Regular Hours	Regular Gross Paid	OT Hours	Total OT Paid	Tota Othe Pa
0	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	BEREZIN	MIKHAIL	NaN	08/10/2015	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0
1	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	GEAGER	VERONICA	M	09/12/2016	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0
2	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	RAMANI	SHRADDHA	NaN	02/22/2016	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0
3	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	ROTTA	JONATHAN	D	09/16/2013	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0
4	2020	17.0	OFFICE OF EMERGENCY MANAGEMENT	WILSON II	ROBERT	P	04/30/2018	BROOKLYN	EMERGENCY PREPAREDNESS MANAGER	ACTIVE	86005.0	per Annum	1820.0	84698.21	0.0	0.0	0

LINEAR

- R^2 score of 0.9398, the lowest among the models tested
- RMSE of 11,093.56, model's predictions are off from the actual salary by about $11,000
- Key features: "Job Title" and "Pay Basis" are the top influential factors in predicting salary
- Effectively captures salary patterns from the data and makes relatively accurate predictions

KNN

- Accurately predicts salaries that clearly fall within a certain class
- Struggles more with salaries that lie near the boundary between two classes
- Key features: "Regular Gross Paid" and "Regular Hours" are the most important features for predicting salary
- Defines more complex salary patterns from the data

DECISION TREE

- R^2 score of 0.967, explaining over 96.7% of salary variance.
- RMSE: 8,220.36.
- Key Features: "Pay Basis_per Annum", "Regular Gross Paid", and “Agency Name_DEPT OF ED PEDAGOGICAL” as the most influential factors in predicting salary.
- Provides a highly interpretable and transparent model structure.

XGBOOST

- R^2 score of 0.9756, the highest among all models tested.
- RMSE: 7,070.64.
- Key Features: "Pay Basis_per Annum", “Agency Name_DEPT OF ED PEDAGOGICAL”, “Agency Name_DEPT OF ED PARA PROFESSIONALS” as primary salary drivers.
- Superior predictive accuracy through sequential error correction.

CONCLUSION

- Best model for accurately predicting salary: XGBoost model
- Best model for understanding which factors have the greatest impact on the salary: Linear Regression model
- “Pay basis” is an influential factor in predicting pay
- Next steps:
  - Examine the effect of additional job-related factors
  - Evaluate how varying certain features influences predicted salaries