NIH Grant Activity Classifier

Erica Zhao, Crystal Kao

What is the NIH?

- Under United States Department of Health and Human Services (HHS)
  - *"NIH's mission is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability."*

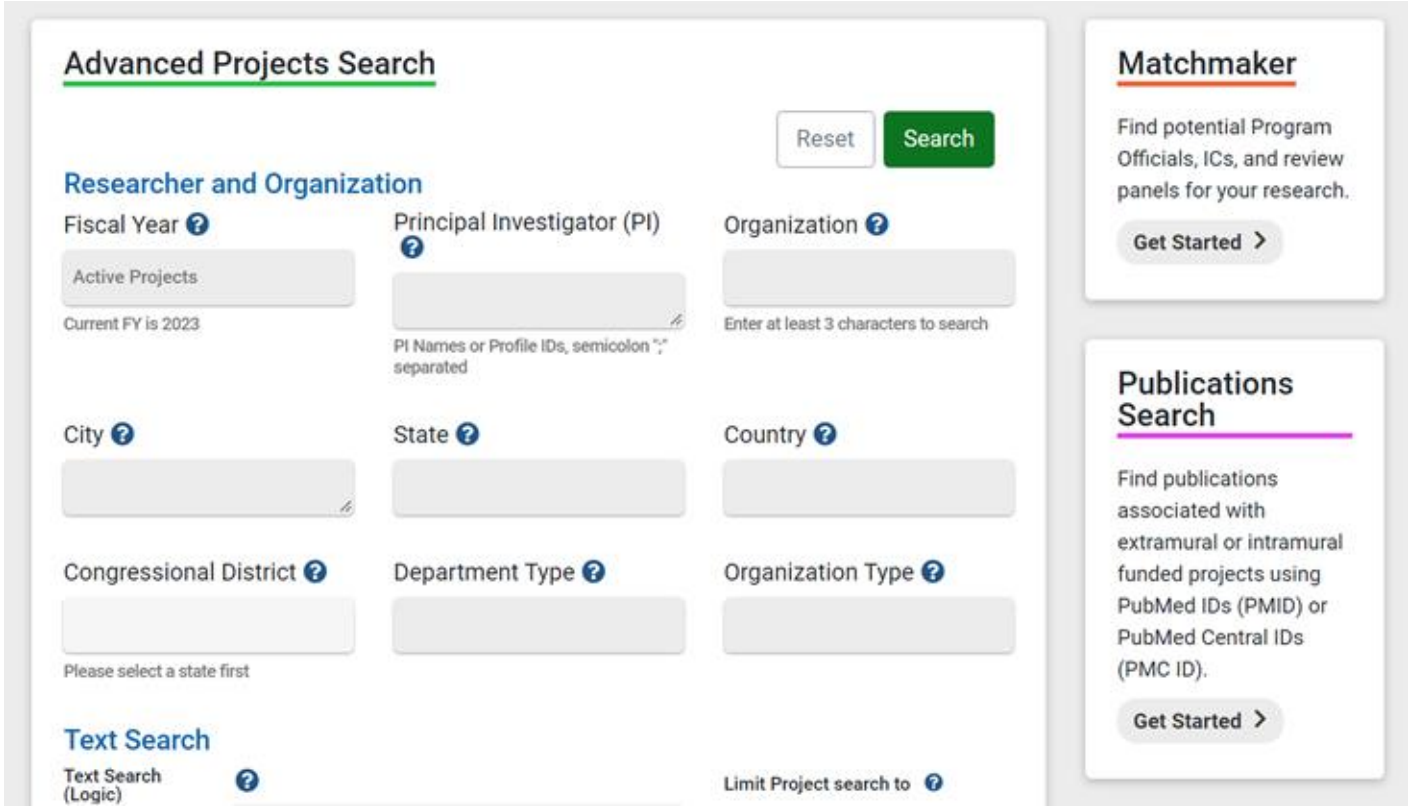
- The NIH consists of 27 institutes and centers (ICs) that each focus on a specific scientific area or function
  - Budget of $48 billion in FY23
  - 83% of NIH funding is awarded to extramural research

Analysis Objective

- The goal of our analysis is to create a model that can categorize the type of grant based on cost and grant length fields
  - The main activity codes of interest are the following 4 because they are the most common pipeline for a PI's (Primary Investigator) development:
    - F: Fellowship grants (for pre- and post-doctorates)
    - T: Research Training grants (also pre- and post-doctorates)
    - K: Career Development Awards (early career)
    - R: Research Grants (for established PIs)
- This analysis may be useful for POs (Program Officer) and data analysts within the NIH to predict what type of grants are coming in and help evaluate the career pipeline

Data Source

- NIH RePORTER (Research Portfolio Online Reporting Tools Expenditures and Results): https://reporter.nih.gov/



Cleaning the Data

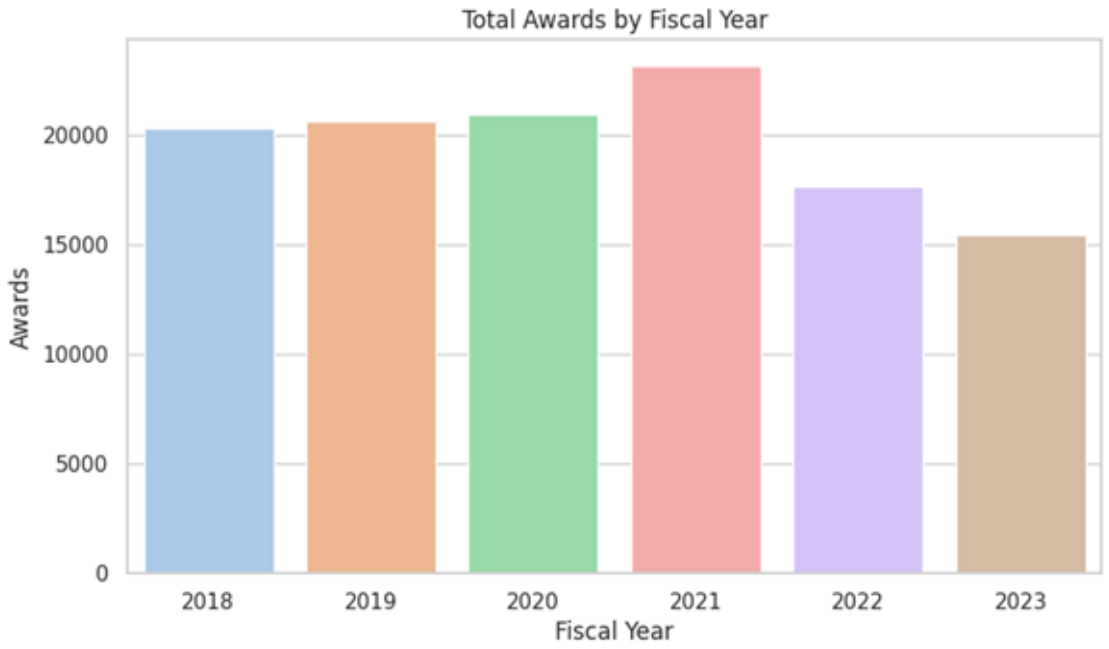
- Base: Adapted from p2_c2_s5_tree_based_models case study.
- Loading Data: Imported FY18-FY23 from NIH RePORTER (118,355 rows)
- Creating Target Variable: Derived from "Activity" codes for grant categorization.
- Filtering: Focused on grant types crucial to analysis - R, K, F, and T series.
- Splitting Data: Allocated 80% for training, 20% for validation (test file is separate).
- Feature Alignment: Removed non-uniform features across datasets.
- Pruning Columns: Excluded non-essential data fields from analysis (project title, contact PI, etc.).
- Handling Missing Data: Imputed using field mean to maintain data integrity.

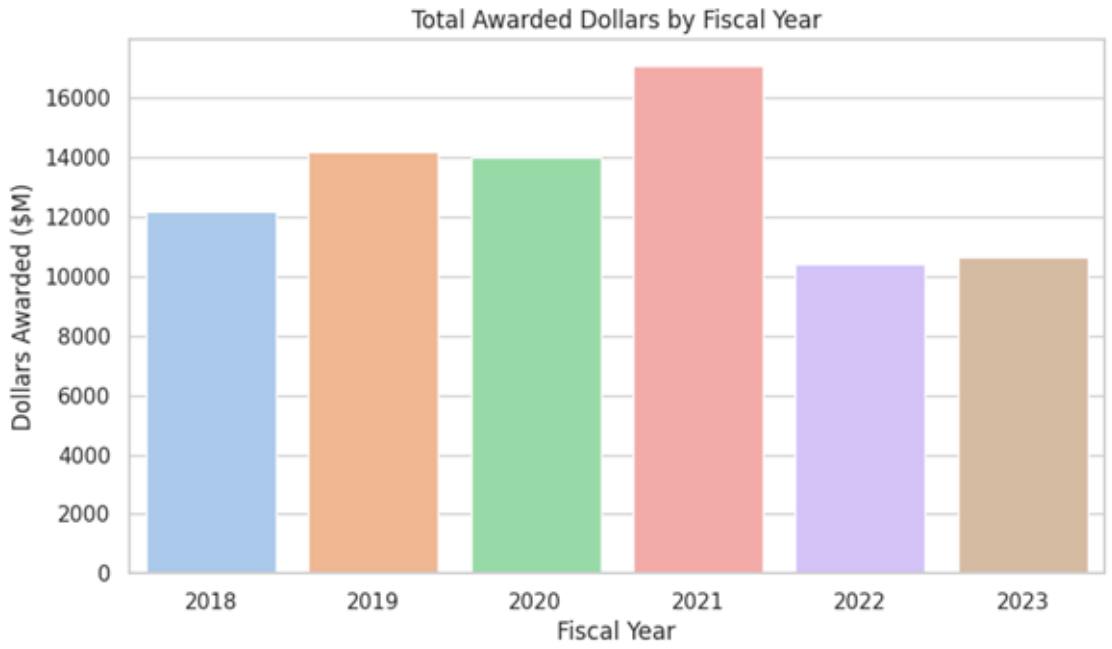
NIH.Spe	Project.	Public.H	Admini	Applica	FOA	Project.	Type	Activity	IC	Contact	Contact	Other.F	Departi	Organiz	Organiz	Organiz	Organiz	Organiz
No NIH Ca	Adipocyte	PROJECT	NIDDK	10464471	PA-21-048	1F32DK13		1 F32	DK	12547591	A, MU	Not Applic	Unavailab	1464901	DANA-FAF	BOSTON	MA	Independ
No NIH Ca	3-Dimensi	NARRATI	NIDCR	10482547	PA-21-259	1R44DE03		1 R44	DE	15409769	AALAMIFA	SEIFABADI	Unavailab	10049265	PEDIAMET	Rockville	MD	Domestic
No NIH Ca	Address;A	PROJECT	NIDA	10590303	RFA-DA-2	1R61DA05		1 R61	DA	7006225	AALSMA, I	RAY, BRAI	PEDIATRIC	577806	INDIANA U	INDIANAP	IN	SCHOOLS
No NIH Ca	Acute;Add	PROJECT	NICHD	10371307	PA-20-206	1K23HD10		1 K23	HD	11863474	AARON, R	Not Applic	PHYSICAL	4134401	JOHNS HO	BALTIMOF	MD	SCHOOLS
No NIH Ca	Acute;Affe	PROJECT	NICHD	10535116	PAR-18-88	1R21HD1C		1 R21	HD	14574456	ABACI TUF	Not Applic	Unavailab	1504801	BOSTON C	BOSTON	MA	Independ
No NIH Ca	Animals;B	n/a	NIGMS	10363324	PA-16-160	4R37GM0		4 R37	GM	7604113	ABALLAY,	Not Applic	Unavailab	6297007	OREGON H	PORTLANI	OR	Domestic
No NIH Ca	Accountin	OTHER	NIA	10522582	PA-20-185	1R01AG07		1 R01	AG	10133980	ABALUCK,	Not Applic	Unavailab	1589901	NATIONAL	CAMBRIDI	MA	Research
No NIH Ca	Anatomy;,	PROJECT	NIMH	10505417	RFA-MH-2	1RF1MH1		1 RF1	MH	16549441	ABBASI AS	Not Applic	NEUROLO	577508	UNIVERSIT	SAN FRAN	CA	SCHOOLS
No NIH Ca	Alloys;Art	PROJECT	NIBIB	10504769	PAR-19-15	1R01EB03		1 R01	EB	12241149	ABBASZAI	WANG, AI	ENGINEER	577510	UNIVERSIT	SANTA CR	CA	BIOMED E
No NIH Ca	3' U	8.	NIGMS	10580284	PAR-21-152	2R15GM1		2 R15	GM	9114482	ABBOTT, A	Not Applic	BIOLOGY	4833601	MARQUET	MILWAUK	WI	SCHOOLS
No NIH Ca	Acute;Add	Electroco	NIMH	10521706	PAR-19-29	1R01MH1		1 R01	MH	10603393	ABBOTT, C	Not Applic	PSYCHIATI	10021612	UNIVERSIT	ALBUQUEI	NM	SCHOOLS
No NIH Ca	Address;A	Projective	NIDCD	10464825	PA-21-051	1F31DC02		1 F31	DC	15589323	ABBOTT, M	Not Applic	OTHER HE	513614	SAN DIEG	SAN DIEG	CA	SCH ALLIE
No NIH Ca	Acute Pair	Project	NINDS	10581160	RFA-TR-22	1DP2NS13		1 DP2	NS	10330774	ABDUS-SA	Not Applic	BIOLOGY	1833202	COLUMBI/	NEW YORI	NY	GRADUAT
No NIH Ca	Advisory C	Project	NIGMS	10428972	PAR-19-34	1K99GM1		1 K99	GM	11915333	ABEBAYEH	Not Applic	BIOMEDIC	1526402	UNIVERSIT	CHARLOTT	VA	BIOMED E
No NIH Ca	Academy;,	PROJECT	NIMHD	10449831	PA-20-190	1K01MD0		1 K01	MD	14837814	ABEBE, EP	Not Applic	NONE	1481402	PURDUE U	WEST LAF	IN	SCHOOLS
No NIH Ca	Abdomina	This	NIDDK	10319759	RFA-DK-20	2U01DK07		2 U01	DK	8264660	ABELL, TH	Not Applic	INTERNAL	4679701	UNIVERSIT	LOUISVILL	KY	SCHOOLS
No NIH Ca	Academic	Public	NIGMS	10412679	PAR-21-14	1T34GM1		1 T34	GM	11165796	ABERNATI	HARRIS, D	CHEMISTR	3499801	HOWARD	WASHING	DC	GRADUAT
No NIH Ca	3-Dimensi	Narrative	NIGMS	10452905	PAR-19-25	1R21GM1		1 R21	GM	10311572	ABHYANK	Not Applic	NONE	7035701	ROCHESTE	ROCHESTE	NY	UNIVERSIT
No NIH Ca	3-Dimensi	Project	NIGMS	10496971	PAR-21-17	1R16GM1		1 R16	GM	10311572	ABHYANK	Not Applic	NONE	7035701	ROCHESTE	ROCHESTE	NY	UNIVERSIT
No NIH Ca	Address;A	PROJECT	NIDA	10494515	PAR-19-28	1R61DA05		1 R61	DA	1901199	ABI-DARG	MOELLER,	PSYCHOLC	5992612	STATE UNI	STONY BR	NY	SCHOOLS
No NIH Ca	Affect;Affi	The	NIA	10370091	PAR-19-07	1R01AG07		1 R01	AG	10569691	ABISAMBF	Not Applic	NEUROSCI	513806	UNIVERSIT	GAINESVII	FL	SCHOOLS
No NIH Ca	Addictive	Project	NIGMS	10412227	PAR-21-16	1R16GM1		1 R16	GM	1968060	ABLORDEI	Not Applic	CHEMISTR	513802	FLORIDA A	TALLAHAS	FL	SCHOOLS
No NIH Ca	Animal M	NARRATI	NINDS	10583656	PA-20-185	1R01NS12		1 R01	NS	7040201	ABOUNAC	Not Applic	MICROBIC	1526402	UNIVERSIT	CHARLOTT	VA	SCHOOLS

Exploratory Visualizations

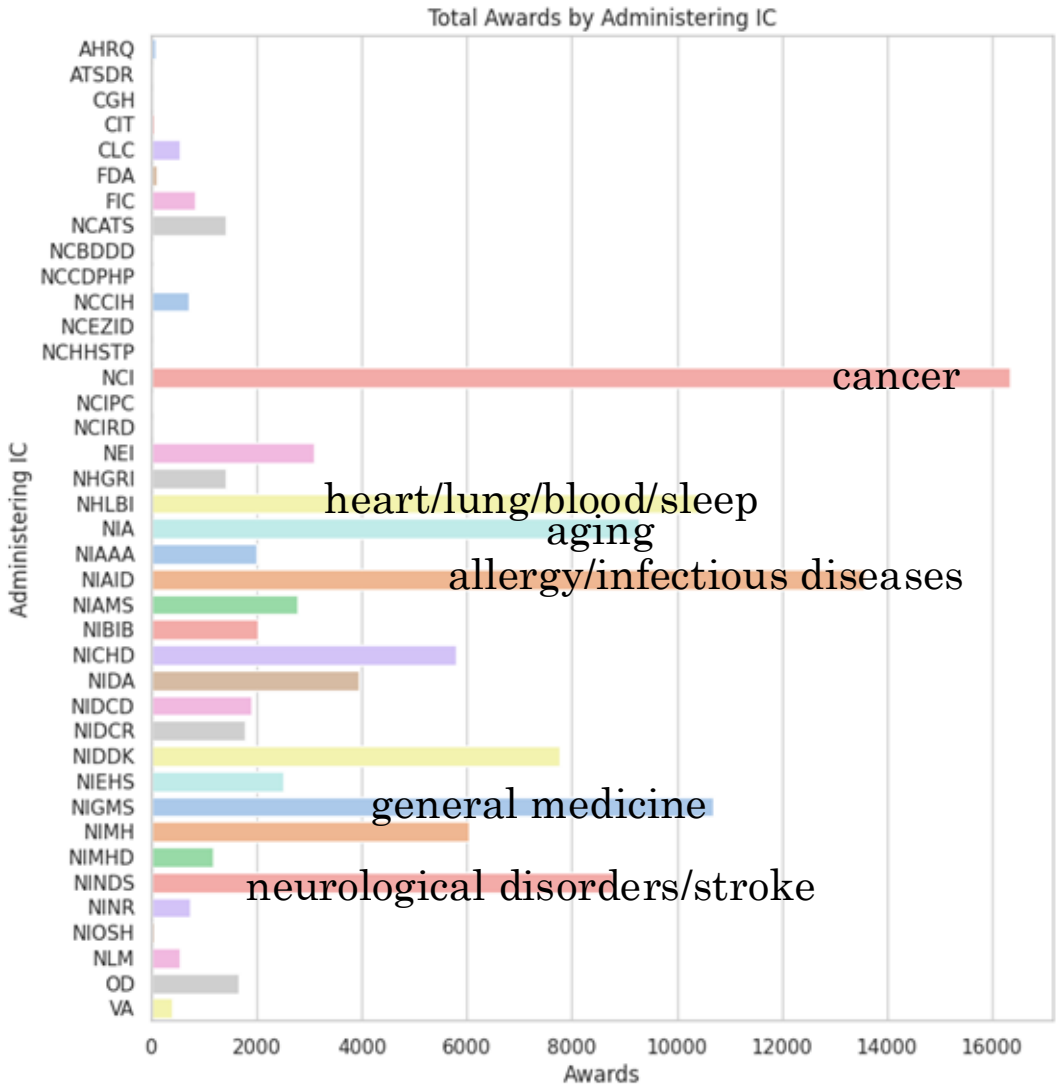
NIH Awards and Funding

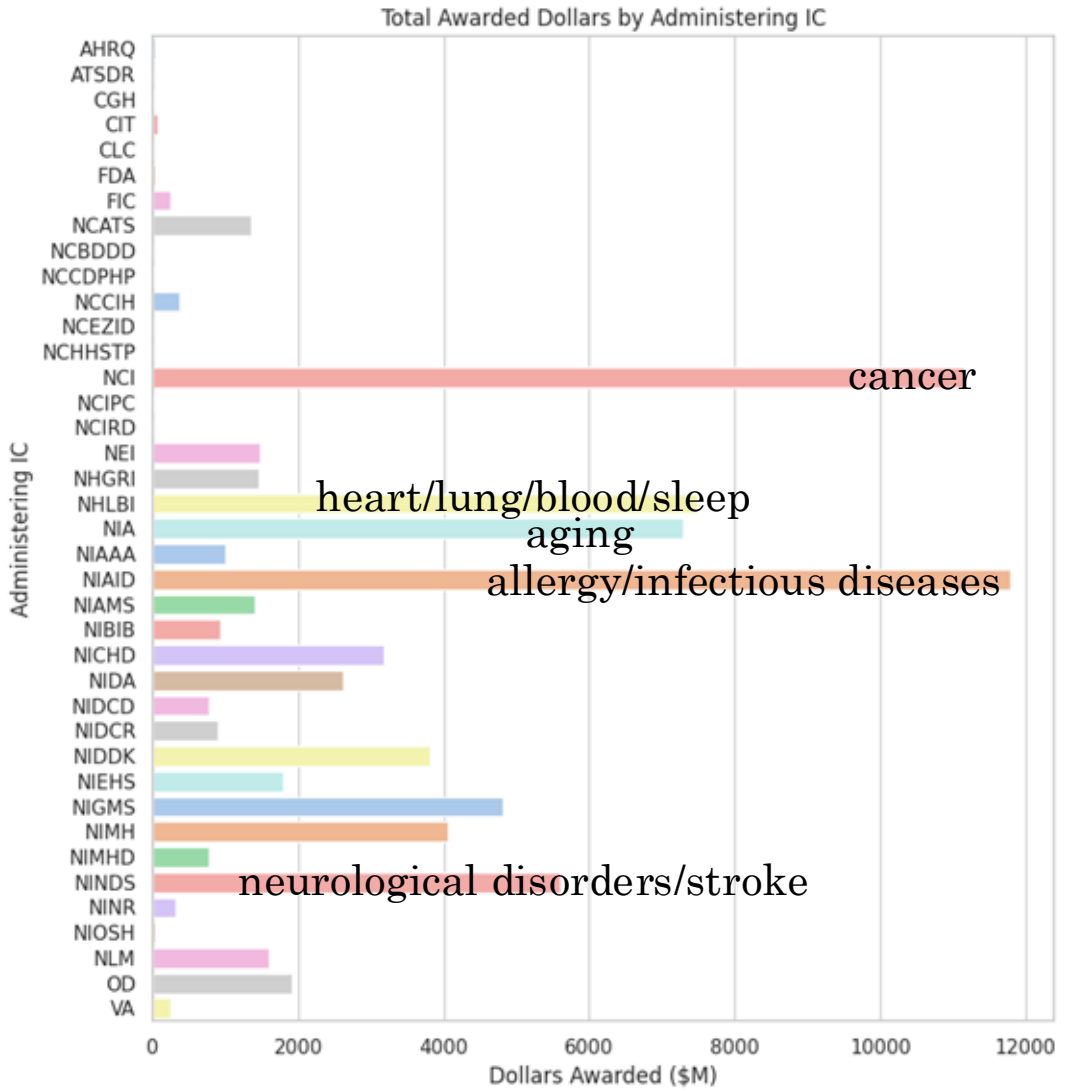
- For the past 6 years, the NIH has awarded ~15,000 - 25,000 grants per FY, with a decrease in the last 2 FYs





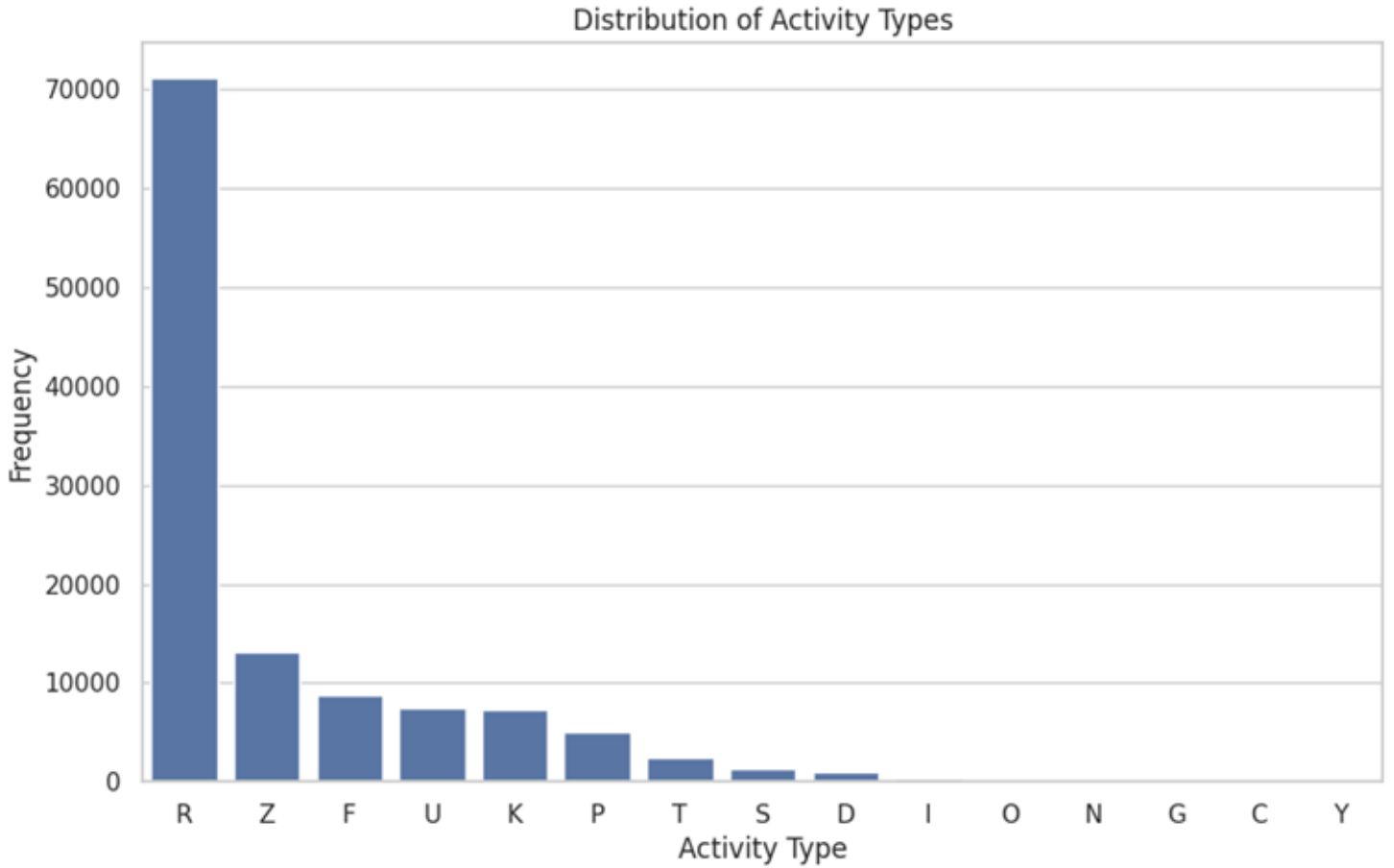
Awards and Funding by IC





Frequency of Activity Types

- Z was excluded to those being intramural (internal to NIH) grants
- F, T, K, and R are the part of the typical PI development pipeline

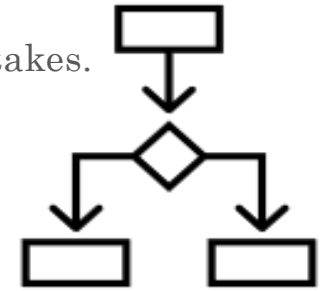


Modeling

Models

- The following 4 models were used in our analysis as our goal is to classify grants into activity types:

1. MLP Classifier
   - Why: Captures complex relationships through its layered structure.
   - Benefit: Adapts well to high-dimensional data, offering nuanced categorization.
2. Logistic Regression
   - Why: Provides a probabilistic approach for binary and multiclass classification.
   - Benefit: Easy to interpret, allows for straightforward assessment of feature impact.
3. Random Forest Classifier
   - Why: Ensemble of decision trees that improve predictive accuracy and control over-fitting.
   - Benefit: Handles categorical data effectively and provides importance scores for features.
4. Histogram-Based Gradient Boosting Classifier
   - Why: Robust to outliers and scalable to large datasets.
   - Benefit: Utilizes gradient boosting to improve prediction as it learns from previous mistakes.



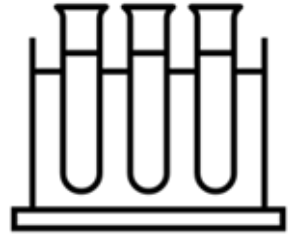
Empirical Results

Model Selection Overview

- **Histogram-Based Gradient Boosting**: Best Score
  - **Score**: 0.92961
  - **Key Parameters**: Learning Rate = 0.1
- **MLP Classifier:** Good Performance
  - **Score**: 0.901421
  - **Key Parameters**: Alpha = 1e-05
- **Decision Tree Classifier**: Close Competitor
  - **Score**: 0.925928
  - **Key Parameters**: Min Samples Leaf = 1
- **Logistic Regression:** Requires Further Tuning
  - **Score**: 0.818733
  - **Key Parameters**: C = 10, Tolerance = 1e-05

	best_score	best_param	best_estimator
0	0.929621	{'model__learning_rate': 0.1, 'model__min_samp...	(HistGradientBoostingClassifier(min_samples_le...
1	0.925928	{'model__min_samples_leaf': 1, 'model__min_sam...	((DecisionTreeClassifier(max_features='sqrt', ...
2	0.901421	{'model__alpha': 1e-05, 'model__learning_rate_...	(MLPClassifier(alpha=1e-05, early_stopping=Tru...
3	0.818733	{'model__C': 10, 'model__tol': 1e-05}	(LogisticRegression(C=10, class_weight='balanc...

Conclusion



Data Preparation: The Foundation of Our Analysis

- Refined Data: Carefully curated from extensive NIH records.
- Targeted Approach: Target variable crafted from 'Activity' codes for precise analysis.
- Validated Methods: Data split and cleaning methods ensure robust model training.
- Missing Data Strategy: Mean imputation preserves data structure, preventing bias.

Impact:

- **Enhanced Clarity**: Categorizes NIH's diverse funding streams with accuracy.
- **Operational Efficiency**: Streamlines grant processing, saving time and resources.
- **Data-Driven Strategy**: Supports informed decision-making within NIH.
- **Transparency**: Allows taxpayers to see how funds are allocated in medical research.

Forward Path:

- Integrate these models for robust, real-time analysis.
- Utilize insights for strategic funding and health policy development.

References

- Final report confirms remdesivir benefits for COVID-19. (2020, October 27). National Institutes of Health (NIH). https://www.nih.gov/news-events/nih-research-matters/final-report-confirms-remdesivir-benefits-covid-19

- What We Do. (n.d.). National Institutes of Health (NIH). https://www.nih.gov/about-nih/what-we-do

- RePORT ⟩ RePORTER. (n.d.). https://reporter.nih.gov/

- U.S. Department of Health and Human Services. (2023, April 11). Types of grant programs. National Institutes of Health. https://grants.nih.gov/grants/funding/funding_program.htm