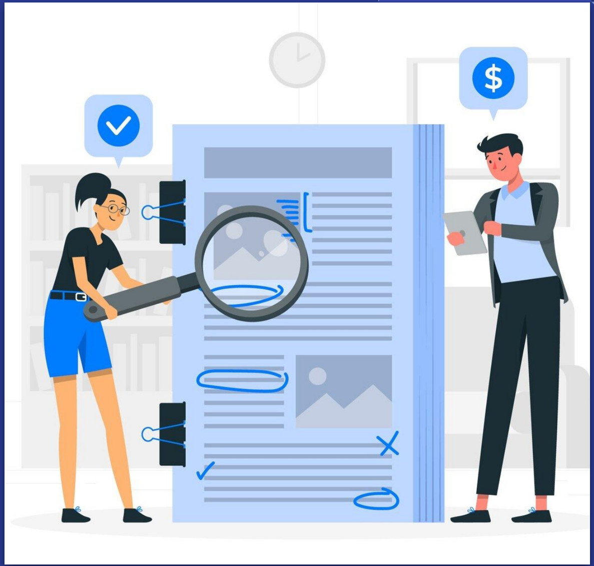
Missing Salary Prediction from Job Postings

Using Machine Learning to Close Information Gaps in the Labor Market

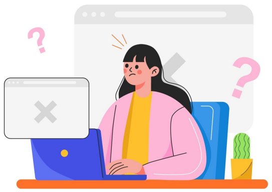


Presentation by - Anushka Dhekne

Agenda

1. Why This Problem Matters?
2. The Real-World Impact & Lightcast Alignment
3. Solution Overview
4. Data Collection & Preparation
5. Feature Engineering & Selection
6. Model Training, Evaluation & Fairness
7. Deployment, Monitoring & Feedback
8. Timeline, Challenges, and Final Thoughts
9. Q&A

Why this problem matters?



- 60% of job listings lack salary data
- Job posts without salary receive 40% fewer applicants
- Salary opacity widens wage gap
- Incomplete workforce & labor insights
- Lack of salary data undermines Lightcast's mission to guide workforce decisions with clarity

What Solving This Unlocks?

Benchmarking without bias

Closing gaps faster with timely data

Inferring market value in real-time

Forecasting skills-to-salary shifts

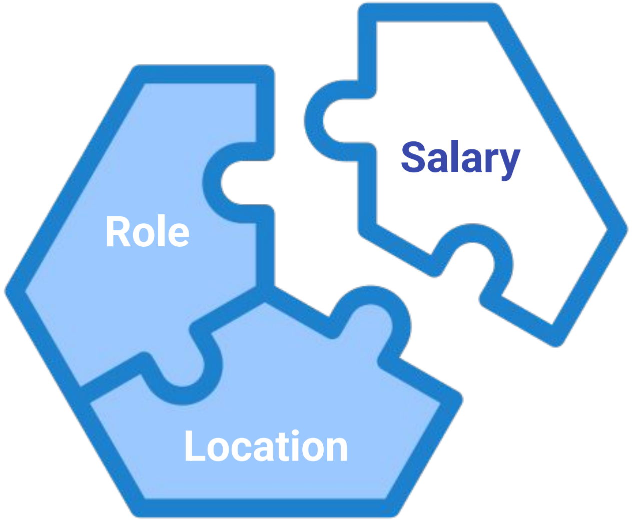
Powering insight-rich dashboards

Why this aligns with Lightcast's Mission?

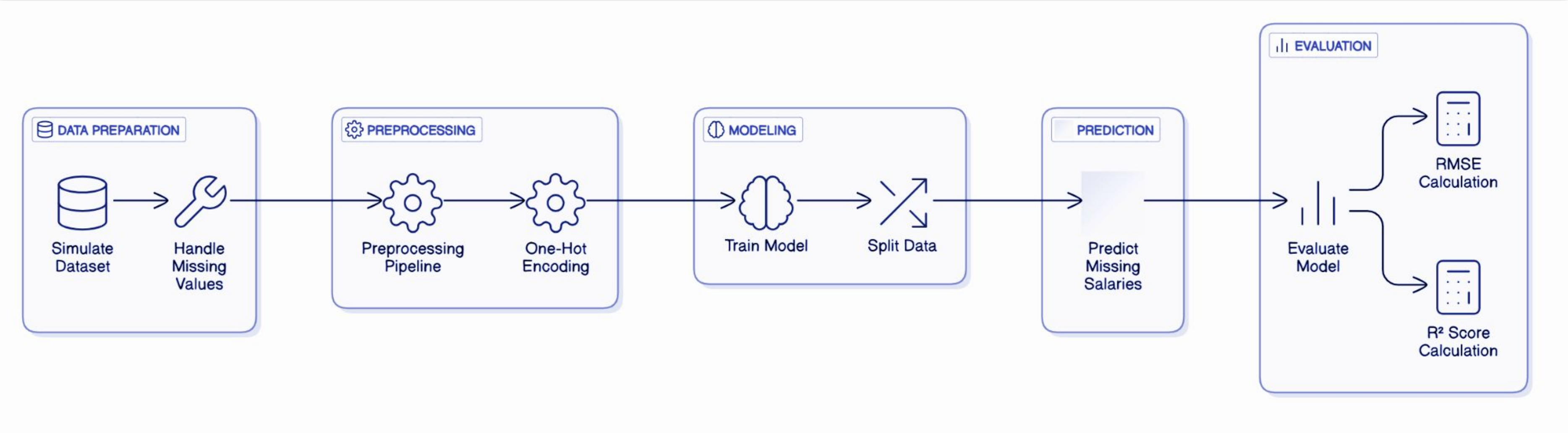
- Supports Lightcast's mission: *"Unlock new possibilities in the labor market"*
- Reveals hidden salary data to drive transparency
- Enables skill-based wage modeling and market clarity
- Informs smarter decisions across business, education, and government

Solution Overview

- Supervised ML model trained on salary-labeled job postings
- Inputs: title, skills, company, location, industry
- Output: predicted salary value or range for missing listings
- Prioritizes explainability and fairness
- Scalable across sectors, geographies, and millions of records



Machine Learning Workflow



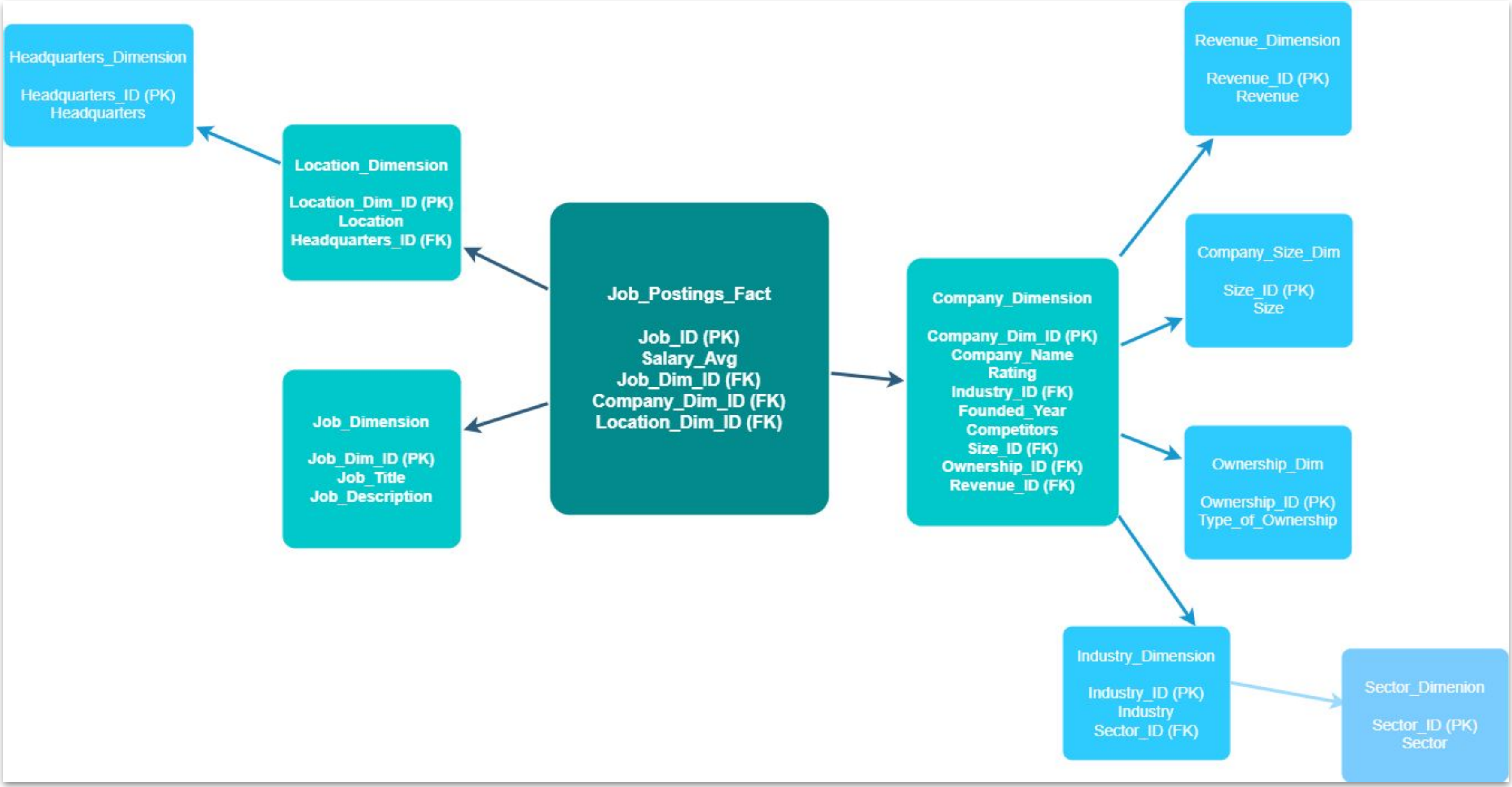
Data Collection & Access

- Structured fields via Snowflake
- Access through secure, scheduled SQL pipelines
- Joinable with external data
- ‘glassdoor_jobs.csv’ with 950+ entries
- ~25% rows had missing salaries (' -1 ')

Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Industry	Sector	Revenue	Competitors
Data Scientist	$53K-$91K (Glassdoor est.)	Data Scientist\nLocation: Albuquerque, NM\nEdu...	3.8	Tecolote Research\n3.8	Albuquerque, NM	Goleta, CA	501 to 1000 employees	1973	Company - Private	Aerospace & Defense	Aerospace & Defense	$50 to $100 million (USD)	-1
Healthcare Data Scientist	$63K-$112K (Glassdoor est.)	What You Will Do:\n\nI. General Summary\n\nThe...	3.4	University of Maryland Medical System\n3.4	Linthicum, MD	Baltimore, MD	10000+ employees	1984	Other Organization	Health Care Services & Hospitals	Health Care	$2 to $5 billion (USD)	-1
Data Scientist	$80K-$90K (Glassdoor est.)	KnowBe4, Inc. is a high growth information sec...	4.8	KnowBe4\n4.8	Clearwater, FL	Clearwater, FL	501 to 1000 employees	2010	Company - Private	Security Services	Business Services	$100 to $500 million (USD)	-1
Data Engineer	-1	Data Engineer\n£50,000 – £70,000 See Advert\n\...	4.5	Anson McCade\n4.5	Kingdom, IL	London, United Kingdom	51 to 200 employees	2000	Company - Private	Staffing & Outsourcing	Business Services	$1 to $5 million (USD)	-1
Business Intelligence Analyst	-1	Business Intelligence Analyst\nAccounting\n50 ...	3.1	Amica Mutual\n3.1	Lincoln, RI	Lincoln, RI	1001 to 5000 employees	1907	Company - Private	Insurance Carriers	Insurance	$1 to $2 billion (USD)	-1

Kaggle Dataset Link

Snowflake Schema Diagram



Data Preparation

- Standardized column names for consistency
- Cleaned & parsed ‘salary_estimate’ to min/max/avg numeric fields
- Split dataset into two parts: for rows with available salary data and for rows with missing or undisclosed salary ('-1')
- Extracted job seniority, job location - city and state
- Applied IQR-based outlier removal

Feature Engineering

- Extracted seniority level from job titles and mapped to ordinal values
- Parsed ‘job_city’ and ‘job_state’ from location, including fallback logic
- Derived ‘posting_age’ from job descriptions and filled missing values
- Applied TF-IDF on job descriptions (top 50 features)
- Created binary flag for missing salaries for modeling tasks

Two-Stage Modeling

Stage 1

Missingness Clarification

- Built Logistic Regression, Random Forest, and KNN models
- Trained on job title, company, seniority, location
- Identified patterns in missing salary data

Stage 2

Salary Prediction

- Random Forest Regressor on known salaries
- Combined structured job features with TF-IDF job description keywords
- Predicted missing salaries & combined into ‘final_salary’ field

Advanced Modeling

- Trained a LightGBM Regressor to predict salaries from clean data
- Applied monotonic constraint so salaries always rise with seniority level
- Feature mix: job title, industry, job state, company, seniority, and posting age
- Cleaned + de-duplicated 100+ features using one-hot encoding
- Result: A powerful, seniority-aware salary model that mimics real-world logic
- Seamless plug-in to existing modular pipeline

Explainability & Fairness

- Treated ‘job_state’ as a geographic fairness lens
- Grouped predictions by state using the Fairlearn library
- Compared average predicted salaries across regions
- Preserved row indices for full-context fairness checks
- Fairness audit shows CA, TX, NC receive highest predicted salaries
- Framework ready to extend to industry, company size, gender or race

Model Output

	job_title	salary_estimate	job_description	rating	company_name	location	headquarters	size	founded	type_of_ownership	industry	sector
0	Data Scientist	-1	JOB CATEGORY:\n\nInformation Services\n\nREQUI...	3.9	Mars\n3.9	Oregon	Mc Lean, VA	10000+ employees	1911	Company - Private	Food & Beverage Manufacturing	Manufacturing
1	Data Scientist	-1	Take your career to new heights working with a...	4.1	Amount\n4.1	Chicago, IL	Chicago, IL	201 to 500 employees	2015	Company - Private	Enterprise Software & Network Solutions	Information Technology
2	Data Science Analyst	-1	Company Overview:\n\nBrightside is an employee...	5.0	Brightside\n5.0	Chandler, AZ	San Francisco, CA	51 to 200 employees	2017	Company - Private	Investment Banking & Asset Management	Finance
3	Data Engineer	-1	Data Engineer\n£50,000 – £70,000 See Advert\n\...	4.5	Anson McCade\n4.5	Kingdom, IL	London, United Kingdom	51 to 200 employees	2000	Company - Private	Staffing & Outsourcing	Business Services
4	Business Intelligence Analyst	-1	Business Intelligence Analyst\nAccounting\n50 ...	3.1	Amica Mutual\n3.1	Lincoln, RI	Lincoln, RI	1001 to 5000 employees	1907	Company - Private	Insurance Carriers	Insurance

BEFORE

	job_title	company_name	industry	job_state	job_city	seniority_level	salary_estimate	final_salary	was_missing
0	Data Scientist	Mars\n3.9	Food & Beverage Manufacturing	Oregon	Oregon	0	-1	114910.0	True
1	Data Scientist	Amount\n4.1	Enterprise Software & Network Solutions	IL	Chicago	0	-1	107440.0	True
2	Data Science Analyst	Brightside\n5.0	Investment Banking & Asset Management	AZ	Chandler	0	-1	105245.0	True
3	Data Engineer	Anson McCade\n4.5	Staffing & Outsourcing	IL	Kingdom	0	-1	86040.0	True
4	Business Intelligence Analyst	Amica Mutual\n3.1	Insurance Carriers	RI	Lincoln	0	-1	72105.0	True

AFTER

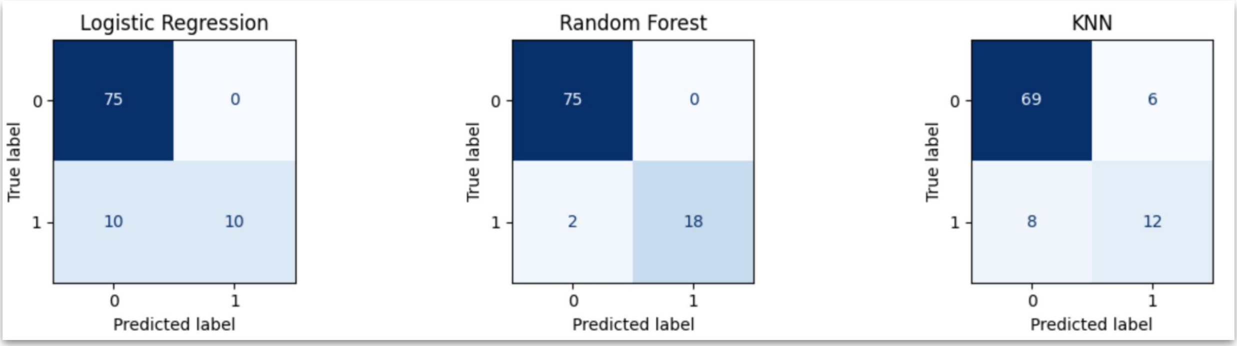
[Interactive Google Sheet]()

Model Evaluation

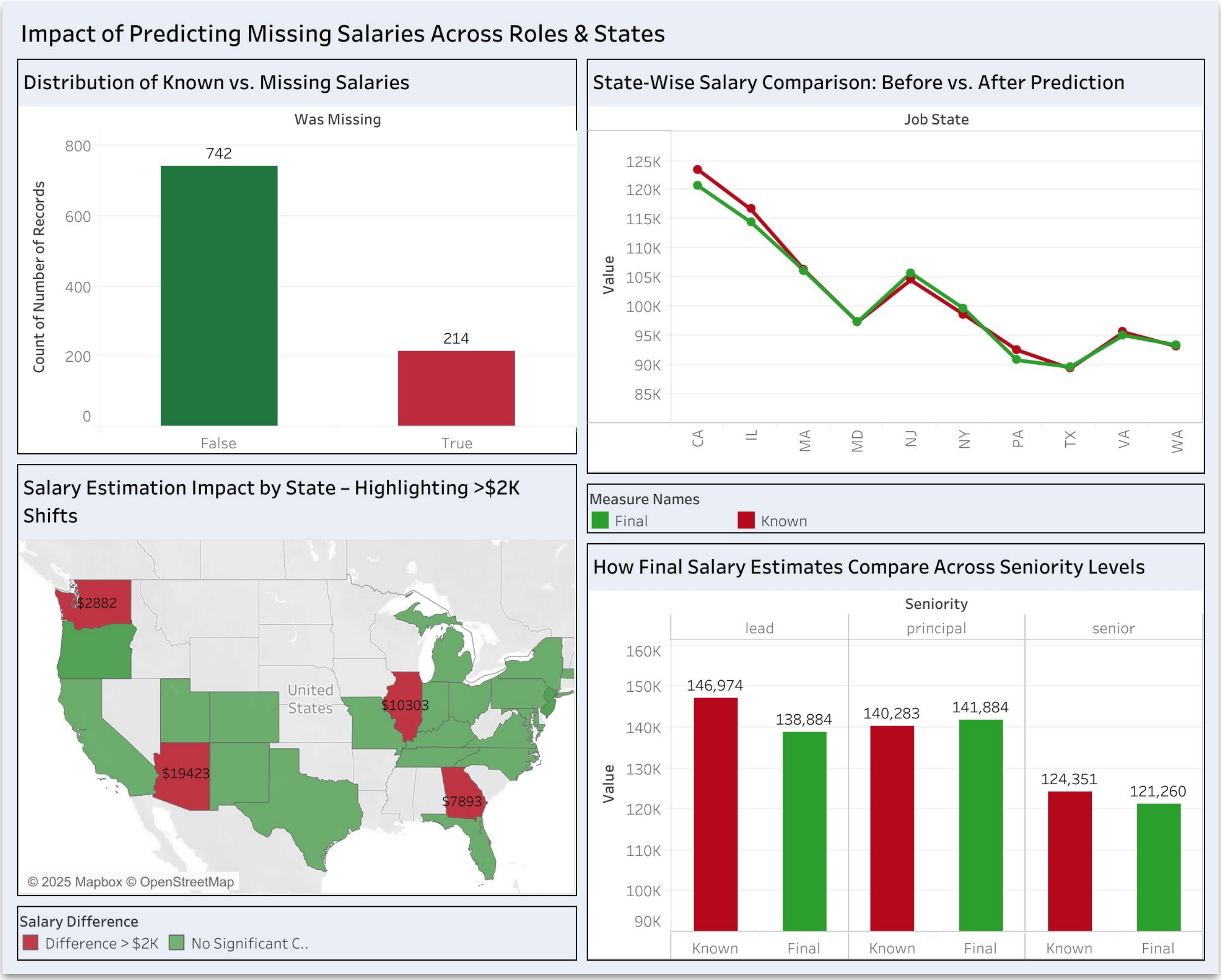
- 97.8% accuracy in detecting missing salaries
- Balanced precision-recall across classifiers
- Random Forest Regressor → MAE: $4,387 | R^2: 0.965
- LightGBM (Monotonic) → MAE: $22,139 | R^2: 0.296
- Confusion matrices show RF outperforms others in true positive detection

```
Model Classification Accuracy
Logistic Regression: $ 0.895
RandomForest Classifier: $ 0.979
KNN: $ 0.853
```

```
Random Forest Regressor Metrics:
MAE: $ 4386.895
RMSE: $ 6704.453
R² Score: $ 0.965
```




```
LightGBM Monotonic Metrics:
MAE: $ 22139.417
RMSE: $ 30052.099
R²: 0.296
```



Dashboard

Tableau Dashboard Link

Source Code

Deployment, Monitoring & Feedback

- Deploy via REST API (FastAPI) or Snowflake batch pipelines
- Monitor prediction drift & model error (MAE, R^2 trends)
- Audit fairness using job_state, industry, company size
- Create feedback loop with hiring teams for real-world corrections
- Retrain model every 3–6 months with updated & flagged data

Project Timeline

Week 1-2	Week 3-4	Week 5-6	Week 7	Week 8	Week 9-10
Data Access & Exploration	Data Cleaning & Feature Engineering	Model Development	Advanced Modeling & Fairness	Evaluation & Business Review	Deployment & Monitoring

Trade-Offs & Challenges

- Accuracy vs Explainability → Prioritized tree models with interpretable outputs
- Generalization → Trained on diverse sectors; room for vertical-specific tuning
- Fairness Audits → Used ‘job_state’ as a proxy; expandable to other dimensions
- Salary Drift → Requires periodic retraining & monitoring
- Modular Design → Enables fast adaptation to data or business changes

Final Takeaways & Strategic Impact

- Built for fairness, accuracy, and interpretability
- Predicts salary gaps with confidence and logic
- Modular pipeline with scalable architecture
- Auditable across sectors, states, and companies
- Deeply aligned with Lightcast's mission

"More than a model - a blueprint for ethical, data-driven labor market insights."

THANK YOU

Q&A