Dataiku Case Study

Income Classification Task

- Milind Bhatnagar

Agenda

01 **Problem Definition**

02 Data Insights

03 Modeling Pipeline

04 Model Evaluations

05 Model Insights

Predicting if a person is making more or less than $50,000 per year and identifying the associated characteristics

US Census Bureau data

Demographic
(age, family, race, etc.)

Economic
(wage, occupation, industry, etc.)

Agenda

01 Problem Definition

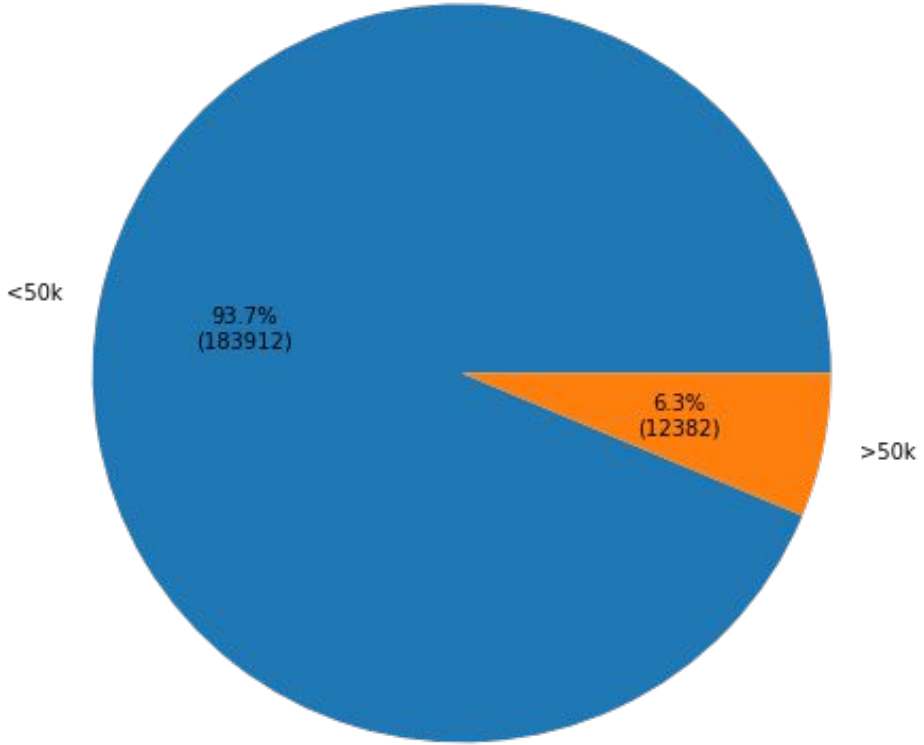
02 **Data Insights**

03 Modeling Pipeline

04 Model Evaluations

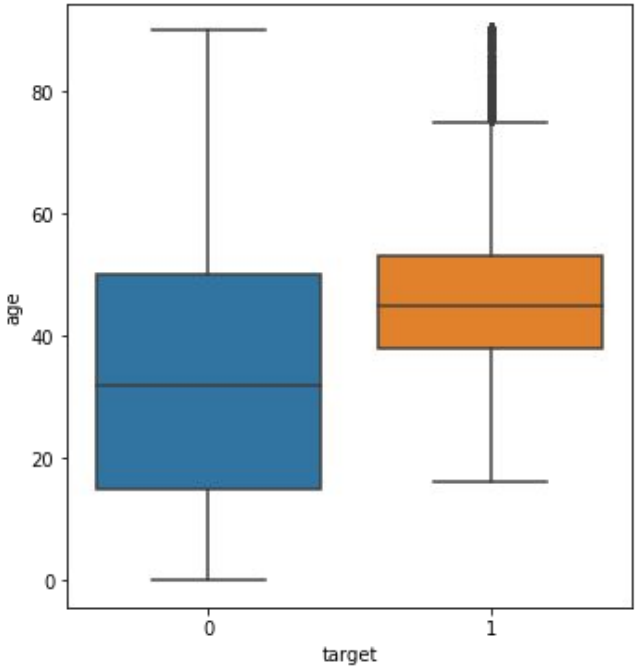
05 Model Insights

Distribution of ~200k people in the training data



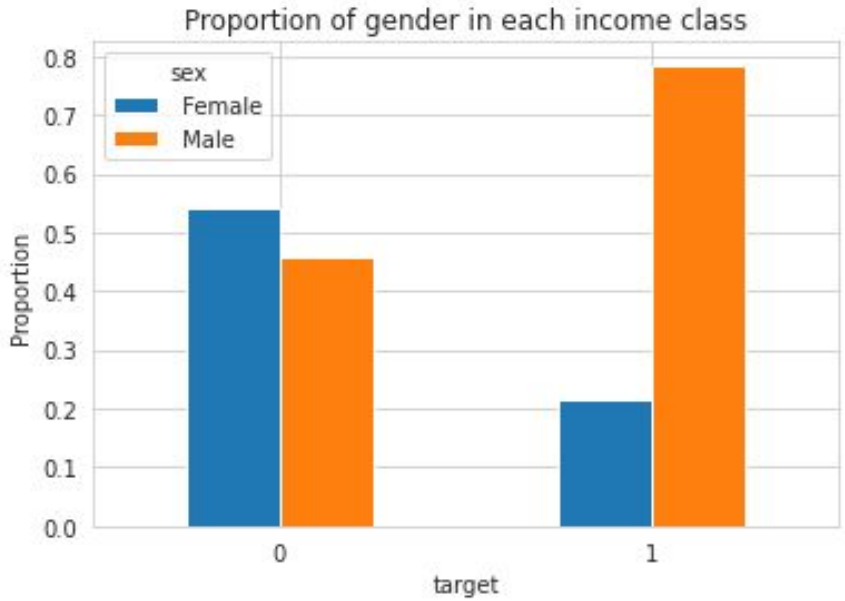
Highly imbalanced classes

Age range for income classes shows minimum working age is 14 and >50k income class has higher concentration of more experienced people



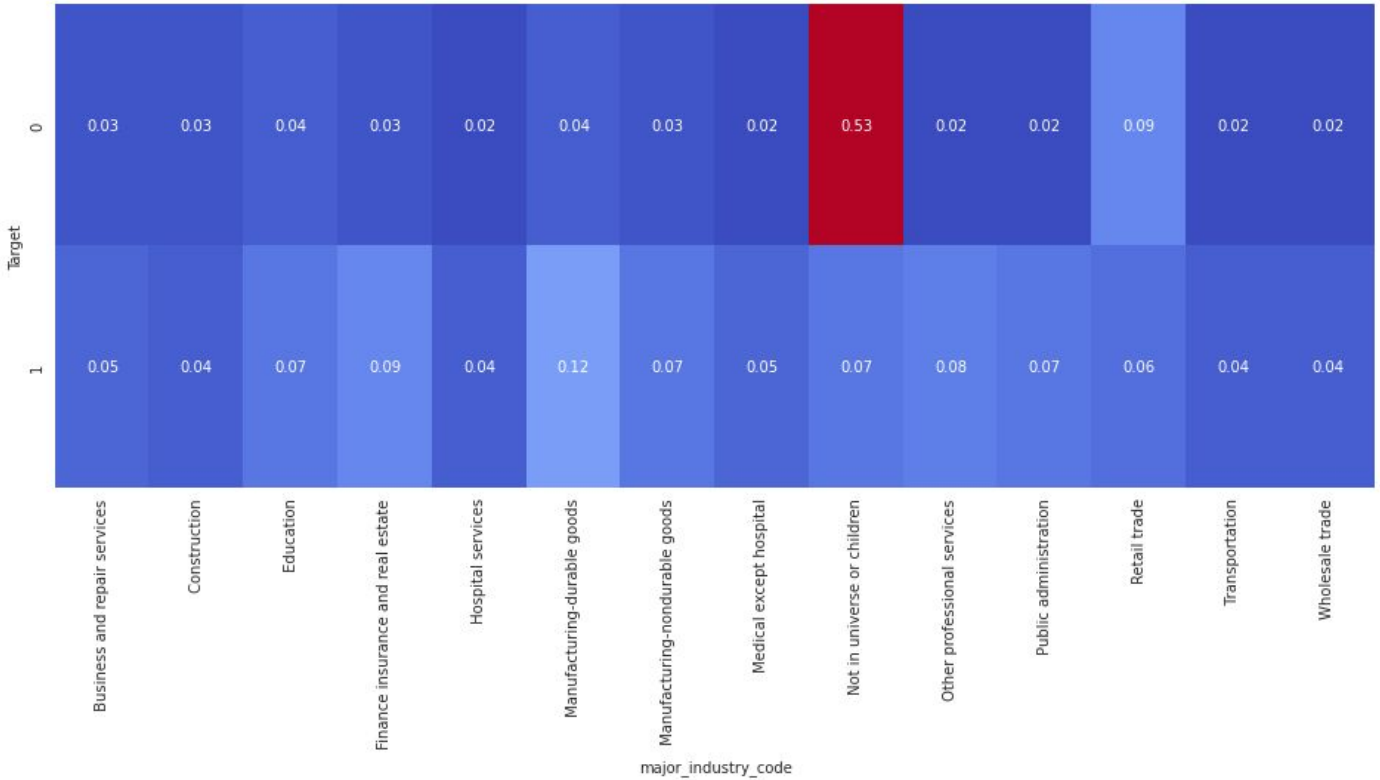
Target- 0: <50k, 1: >50k

Higher proportion of men are in the income class >50k



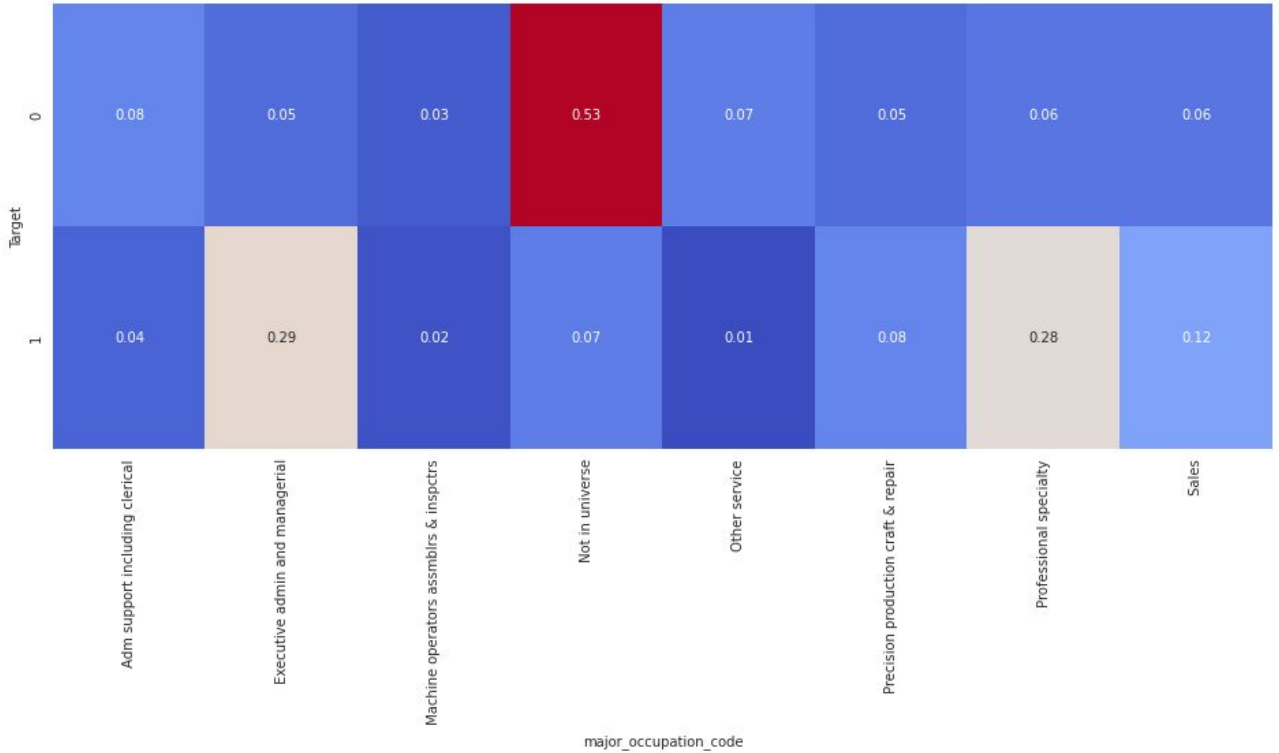
Target- 0: <50k, 1: >50k

A significantly greater proportion of people with ‘Not in universe or children’ have income below 50k



Target- 0: <50k, 1: >50k

Similar observation is seen for major occupation code, additionally executive admin and professional specialty have higher proportion to be in income class >50k



Target- 0: <50k, 1: >50k

Agenda

01 Problem Definition

02 Data Insights

03 **Modeling Pipeline**

04 Model Evaluations

05 Model Insights

Machine learning pipeline

Machine learning pipeline

- Remove irrelevant features
- Drop duplicates

Machine learning pipeline

- Cube root transformations
- New features:
  - Estimated capital stock profit = Capital Gain - Capital Loss + Dividend from stocks
  - Estimated income = Estimated capital stock profit + Wage per hour * Weeks worked in years

Machine learning pipeline

- Remove highly correlated features
- Perform feature selection using Boruta and random forest
- 24 features selected

Machine learning pipeline

- Baseline (All classified as <50k)
- Light GBM
- CatBoost
- Same models with balanced classes

Agenda

01 Problem Definition

02 Data Insights

03 Modeling Pipeline

04 **Model Evaluations**

05 Model Insights

CatBoost predicts the most accurate income classifications (marginally better than Light GBM)

	Baseline (All classified as <50k)	Light GBM (unbalanced*)	CatBoost (unbalanced*)
Accuracy	0.94	0.96	0.96
F1 Score	-	0.59	0.59
Precision	-	0.75	**0.76**
Recall	-	0.48	0.48

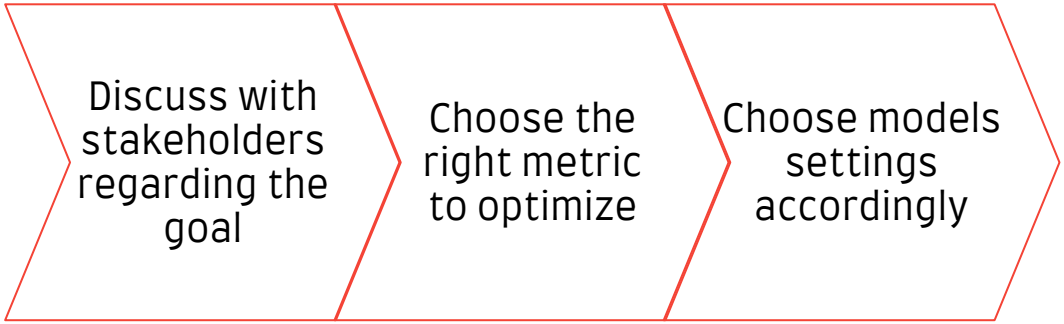
*Class weights not taken into consideration

When class weights are taken, recall increases but other metrics fall

	Baseline (All classified as <50k)	Light GBM (balanced*)	CatBoost (balanced*)
Accuracy	0.94	0.88	0.88
F1 Score	-	0.48	0.48
Precision	-	0.331	0.328
Recall	-	**0.888**	**0.891**

*Class weights taken into consideration

For model selection, it is important to-



Agenda

01 Problem Definition

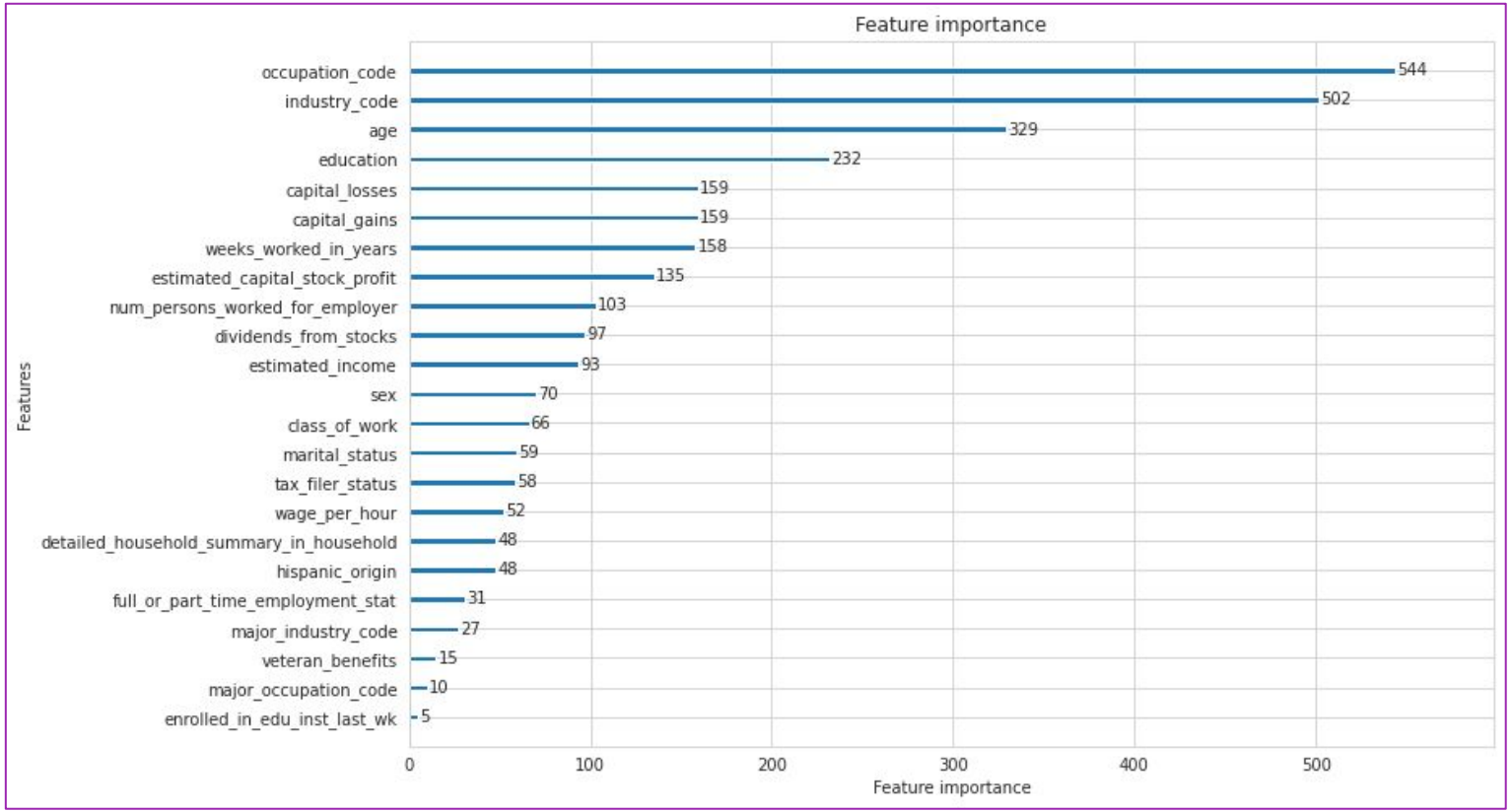
02 Data Insights

03 Modeling Pipeline

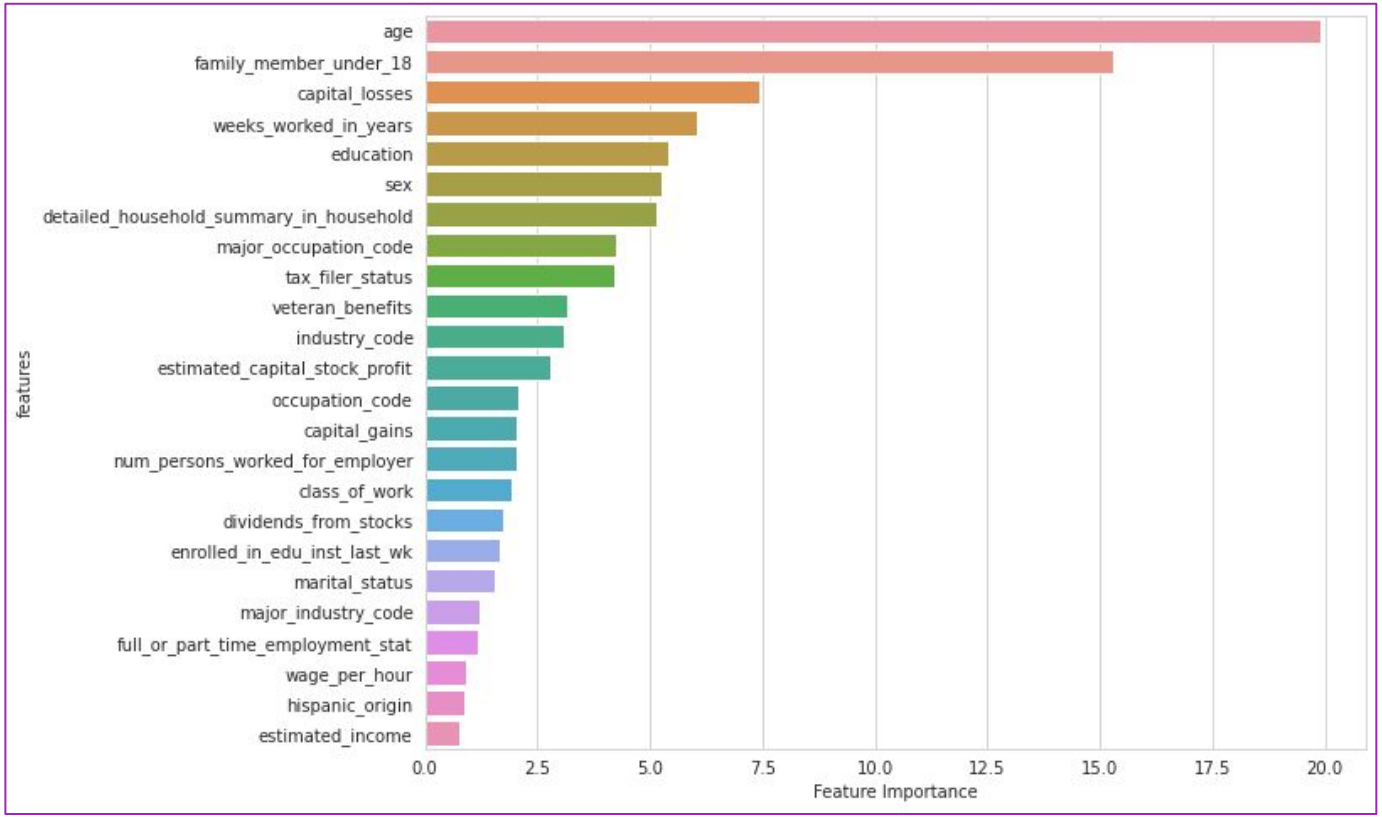
04 Model Evaluations

05 **Model Insights**

Feature importance (Light GBM)



Feature importance (CatBoost)



THANK YOU!

Any questions?