TeleGam

Combining **Visualization** and **Verbalization** for Interpretable Machine Learning

VIS 2019
Vancouver, Canada

Fred Hohman
@fredhohman
Georgia Tech

Arjun Srinivasan
Georgia Tech

Steven Drucker
Microsoft Research

Georgia Tech

Microsoft Research

Google

ai is

- ai is **dangerous**
- ai is **the new electricity**
- ai is **taking over**
- ai is **a crapshoot**
- ai is **overhyped**
- ai is **just if statements**
- ai is **bad**
- ai is **the future**
- ai is **everywhere**
- ai is **scary**

Google Search | I'm Feeling Lucky

Report inappropriate predictions

*While **building and deploying** ML models is now an increasingly common practice, **interpreting** models is not.*

GAMUT

Operationalize

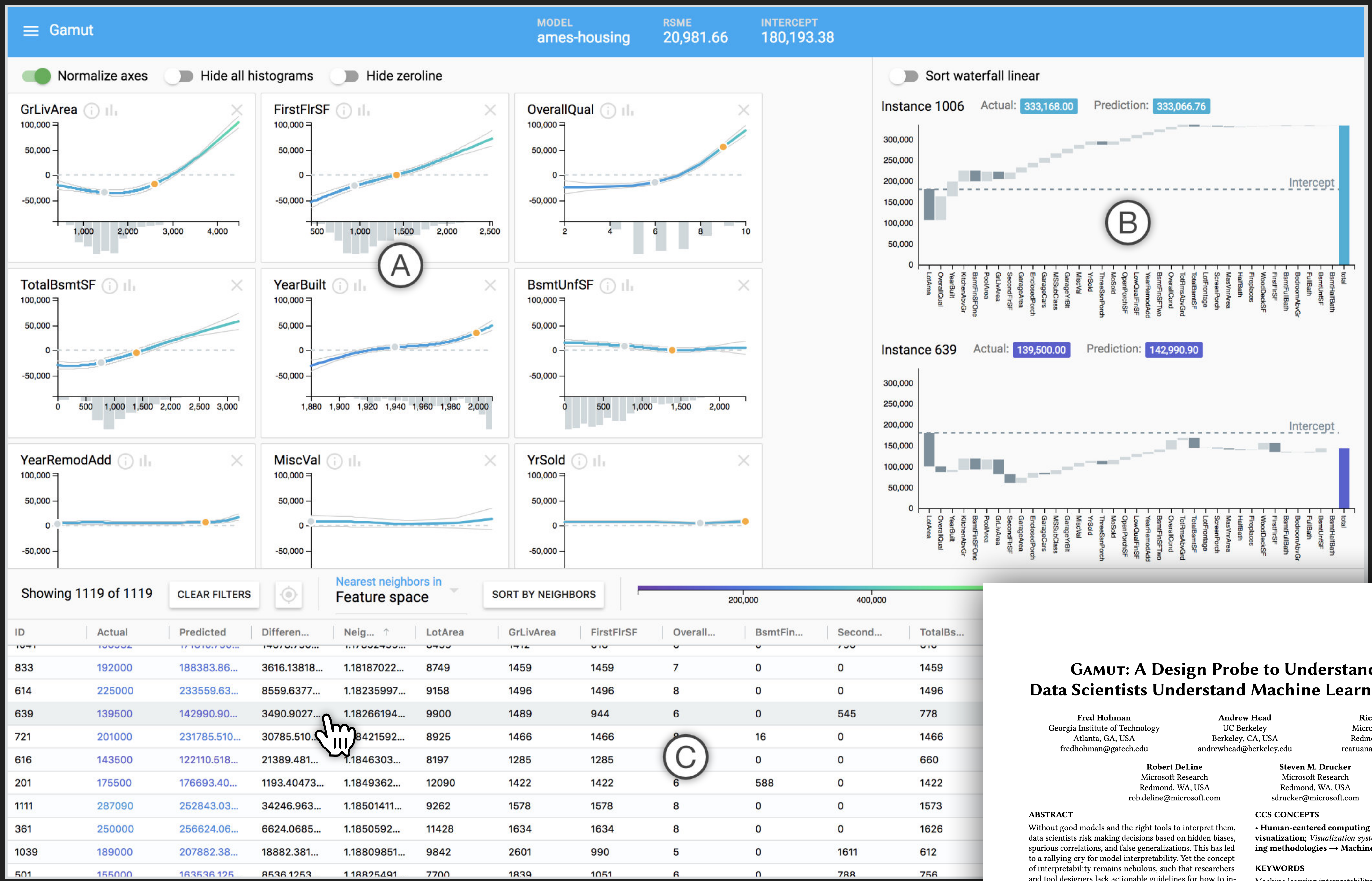
Interpretability in design probe

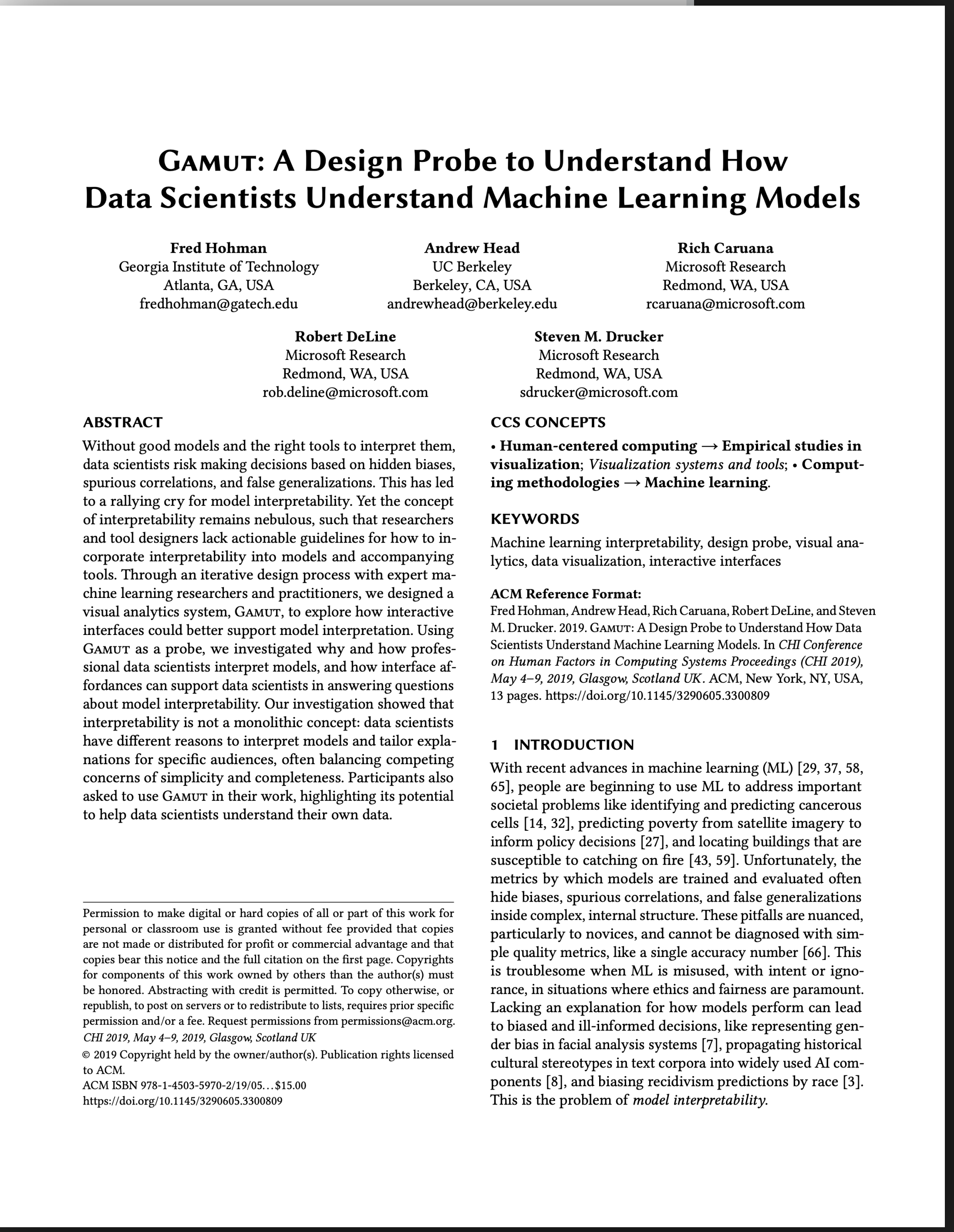
GAMs

Use generalized additive models

Investigation

Of emerging practice of interpretability w/ industry practitioners





GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models

Fred Hohman, Georgia Institute of Technology, Atlanta, GA, USA, fredhohman@gatech.edu
Andrew Head, UC Berkeley, Berkeley, CA, USA, andrewhead@berkeley.edu
Rich Caruana, Microsoft Research, Redmond, WA, USA, rcaruana@microsoft.com
Robert DeLine, Microsoft Research, Redmond, WA, USA, rob.deline@microsoft.com
Steven M. Drucker, Microsoft Research, Redmond, WA, USA, sdrucker@microsoft.com

ABSTRACT

Without good models and the right tools to interpret them, data scientists risk making decisions based on hidden biases, spurious correlations, and false generalizations. This has led to a rallying cry for model interpretability. Yet the concept of interpretability remains nebulous, such that researchers and tool designers lack actionable guidelines for how to incorporate interpretability into models and accompanying tools. Through an iterative design process with expert machine learning researchers and practitioners, we designed a visual analytics system, GAMUT, to explore how interactive interfaces could better support model interpretation. Using GAMUT as a probe, we investigated why and how professional data scientists interpret models, and how interface affordances can support data scientists in answering questions about model interpretability. Our investigation showed that interpretability is not a monolithic concept: data scientists have different reasons to interpret models and tailor explanations for specific audiences, often balancing competing concerns of simplicity and completeness. Participants also asked to use GAMUT in their work, highlighting its potential to help data scientists understand their own data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2019, May 4–9, 2019, Glasgow, Scotland UK
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300809

CCS CONCEPTS

• **Human-centered computing → Empirical studies in visualization**; *Visualization systems and tools*; • **Computing methodologies → Machine learning**.

KEYWORDS

Machine learning interpretability, design probe, visual analytics, data visualization, interactive interfaces

ACM Reference Format:
Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300809

1 INTRODUCTION

With recent advances in machine learning (ML) [29, 37, 58, 65], people are beginning to use ML to address important societal problems like identifying and predicting cancerous cells [14, 32], predicting poverty from satellite imagery to inform policy decisions [27], and locating buildings that are susceptible to catching on fire [43, 59]. Unfortunately, the metrics by which models are trained and evaluated often hide biases, spurious correlations, and false generalizations inside complex, internal structure. These pitfalls are nuanced, particularly to novices, and cannot be diagnosed with simple quality metrics, like a single accuracy number [66]. This is troublesome when ML is misused, with intent or ignorance, in situations where ethics and fairness are paramount. Lacking an explanation for how models perform can lead to biased and ill-informed decisions, like representing gender bias in facial analysis systems [7], propagating historical cultural stereotypes in text corpora into widely used AI components [8], and biasing recidivism predictions by race [3]. This is the problem of *model interpretability*.

GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models.
Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, Steven Drucker. *CHI, 2019.*

Visualization

Explanations

Show model context

Interactive analytics

Rely on user interpretation

Visualization 📊

Explanations

Show model context

Interactive analytics

Rely on user interpretation

📝 Verbalization

Explanations

Direct and concise

Less cognitive load

No training needed

Visualization + Verbalization

Explanations

Explanations

TeleGam

Automatically generate natural language statements, or **verbalizations**, to complement explanatory **visualizations** for machine learning models.

Visualization Explanations + **Verbalization** Explanations

Demo

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude:

Model Feature Summary

Instance Feature Summary Settings

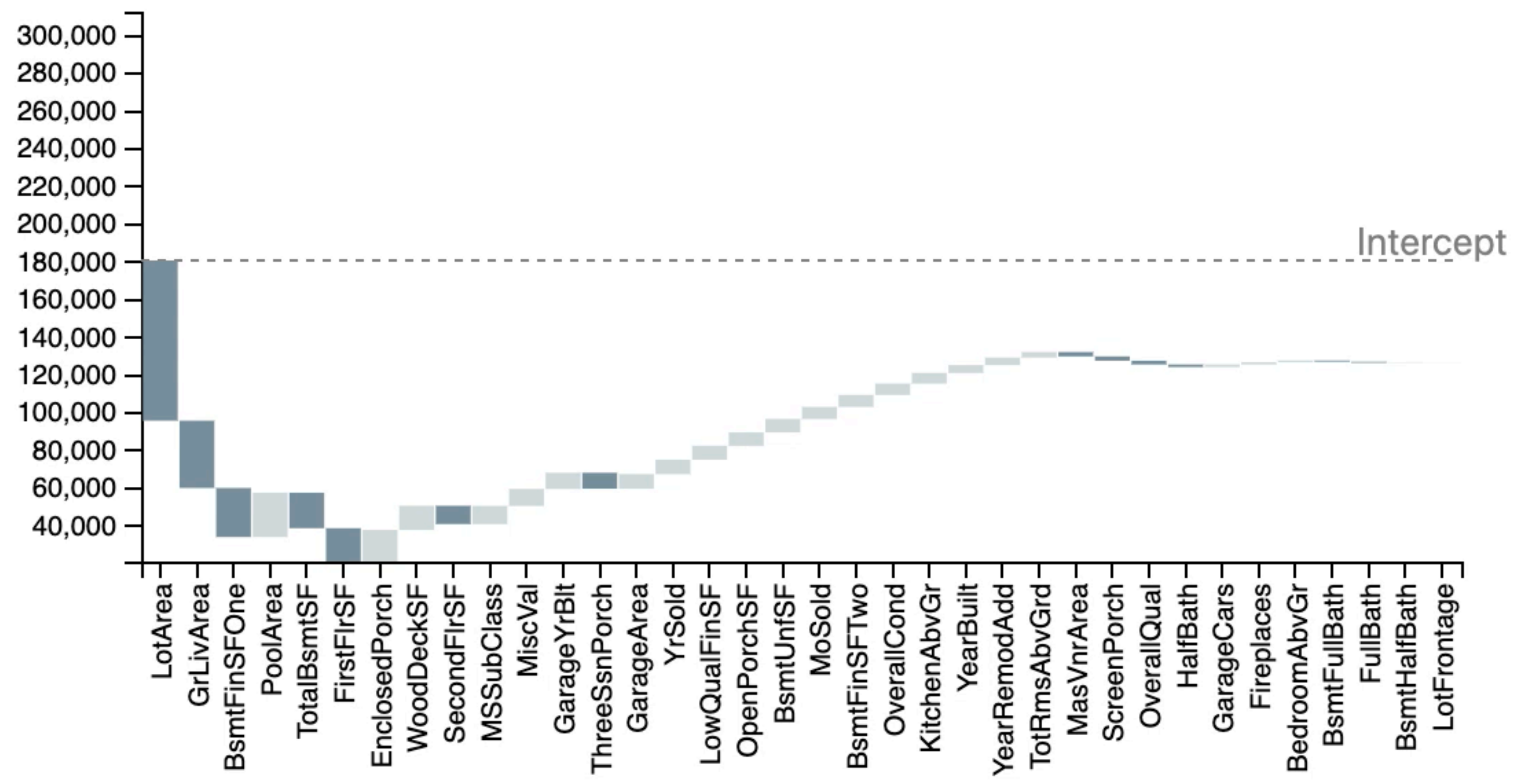
Instance Comparison Summary Settings

Base Instance Summary

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**

Some features have a notable impact on the prediction.



Comparison Summary

Compared Instance Summary

Instance , Actual: , Prediction:

Compared instance:

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude: ☑

Model Feature Summary

Instance Feature Summary Settings

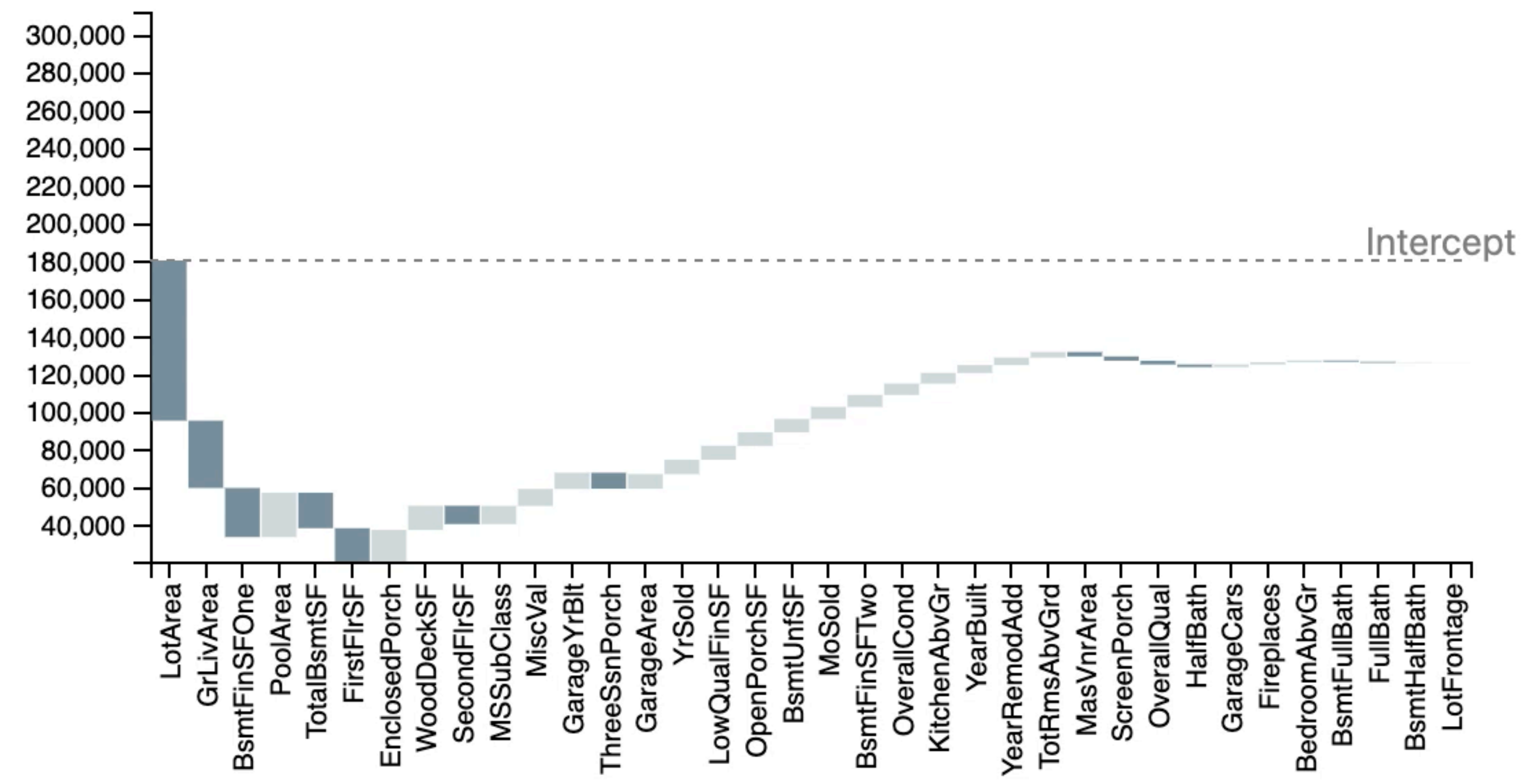
Instance Comparison Summary Settings

Base Instance Summary

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**

Some features have a notable impact on the prediction.



Comparison Summary

Compared Instance Summary

Instance , Actual: , Prediction:

Compared instance:

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude:

Model Feature Summary

Instance Feature Summary Settings

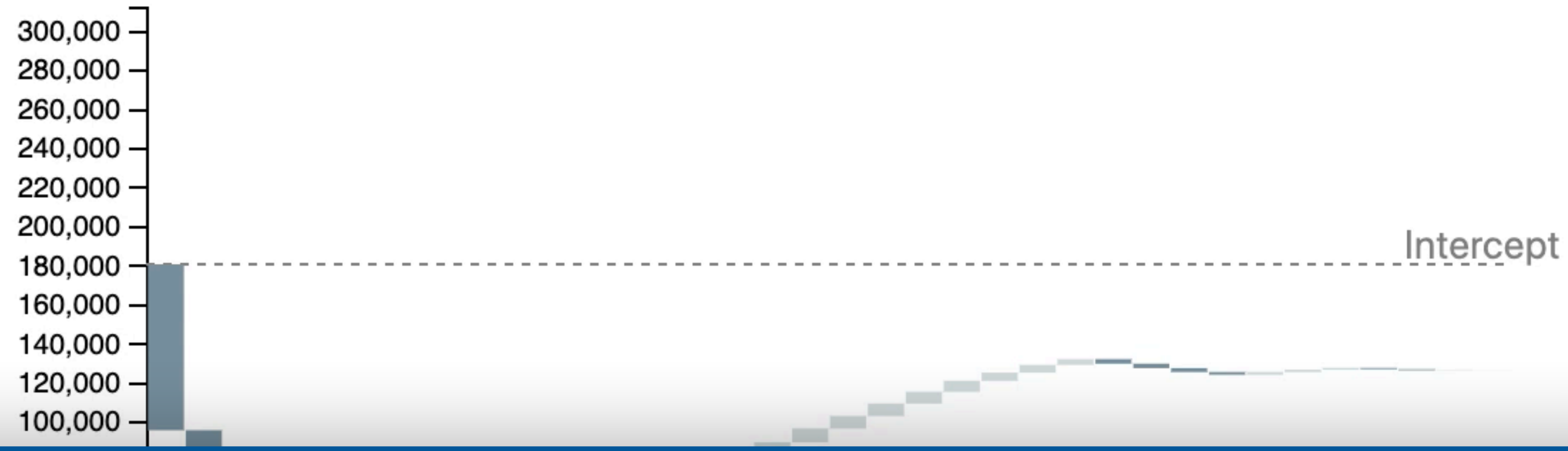
Instance Comparison Summary Settings

Base Instance Summary

Some features have a notable impact on the prediction.

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**



Visualize each feature's global impact on model, grouped by **verbalization**

Comparison Summary

Compared Instance Summary

Instance , Actual: , Prediction:

Compared instance:

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude:

Model Feature Summary

Instance Feature Summary Settings

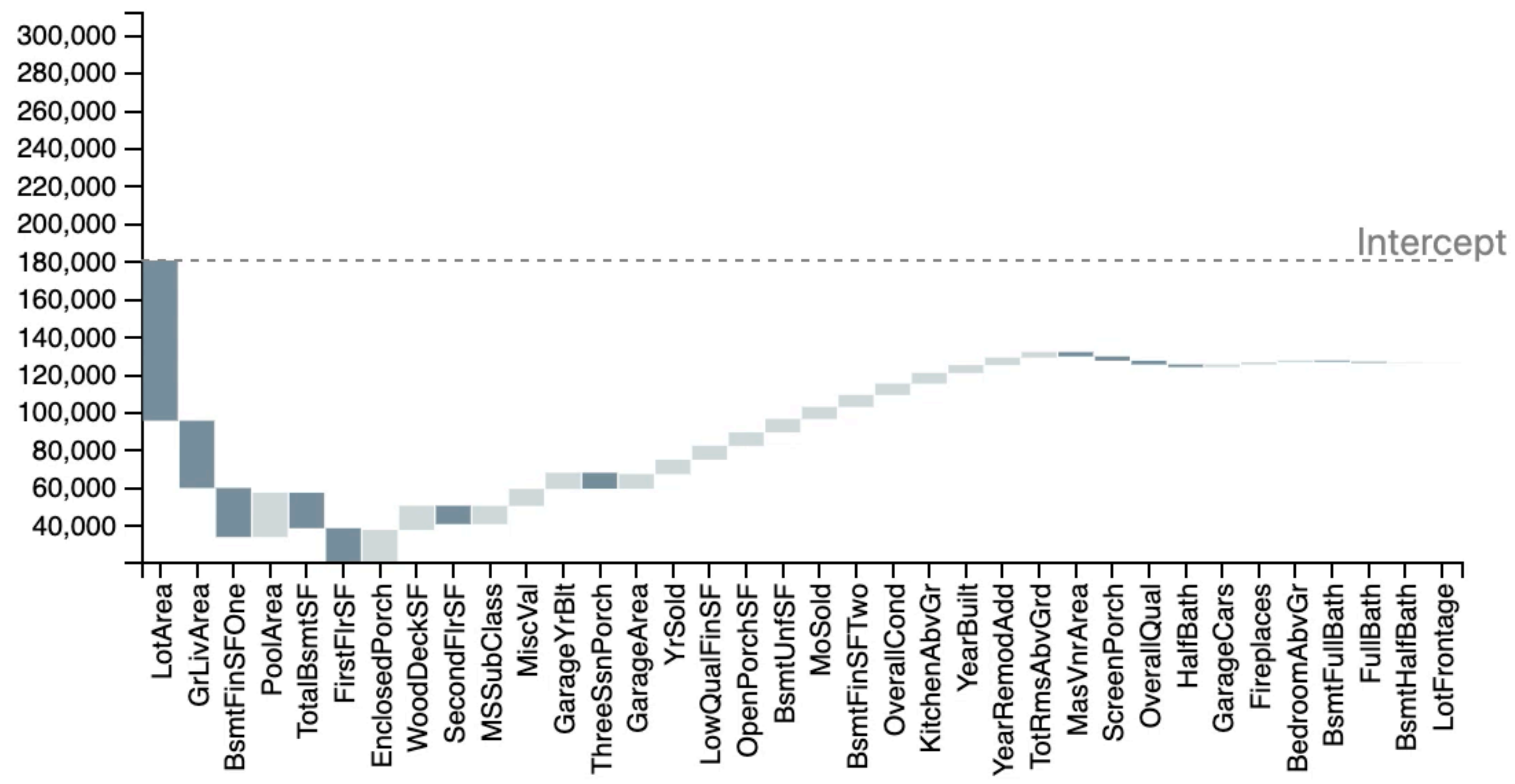
Instance Comparison Summary Settings

Base Instance Summary

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**

Some features have a notable impact on the prediction.



Comparison Summary

Compared Instance Summary

Instance , Actual: , Prediction:

Compared instance:

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude:

Model Feature Summary

Instance Feature Summary Settings

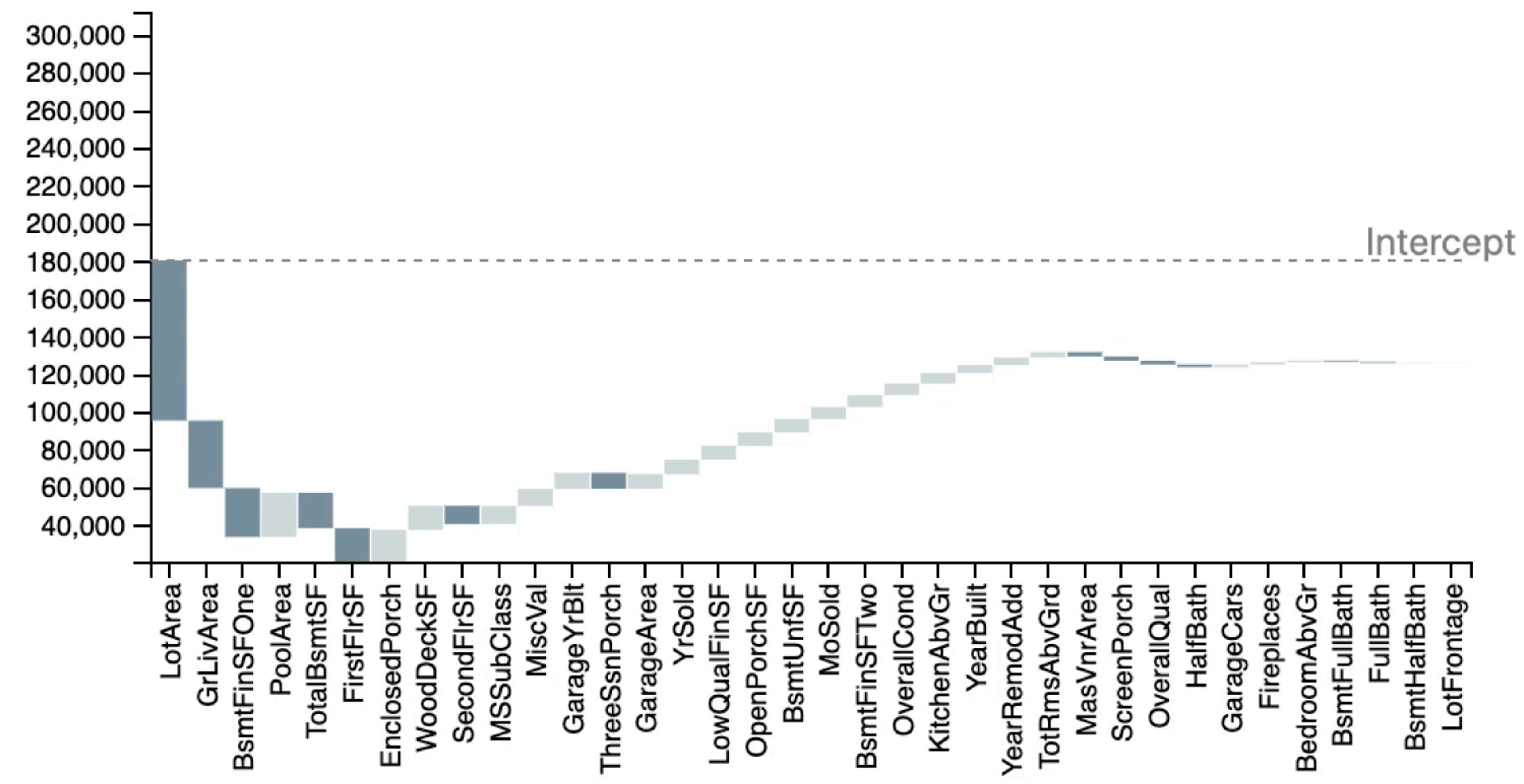
Instance Comparison Summary Settings

Base Instance Summary

Some features have a notable impact on the prediction.

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**



Comparison Summary

Compared Inst

Interactively highlight **verbalization** in context of the **visualization**

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude:

Model Feature Summary

Instance Feature Summary Settings

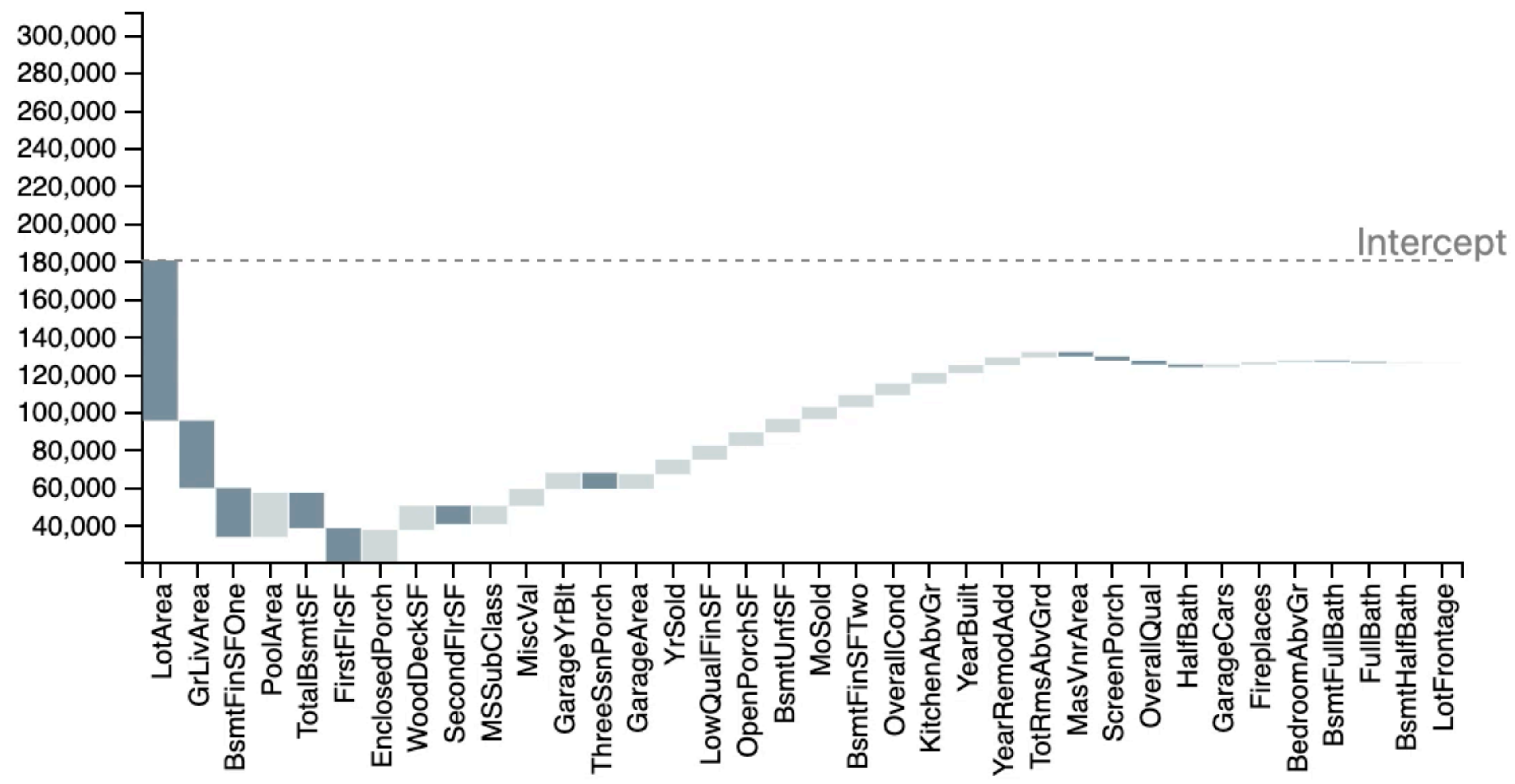
Instance Comparison Summary Settings

Base Instance Summary

Some features have a notable impact on the prediction.

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**



Comparison Summary

Compared Instance Summary

Instance , Actual: , Prediction:

Compared instance:

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude: ☑

Model Feature Summary

Instance Feature Summary Settings

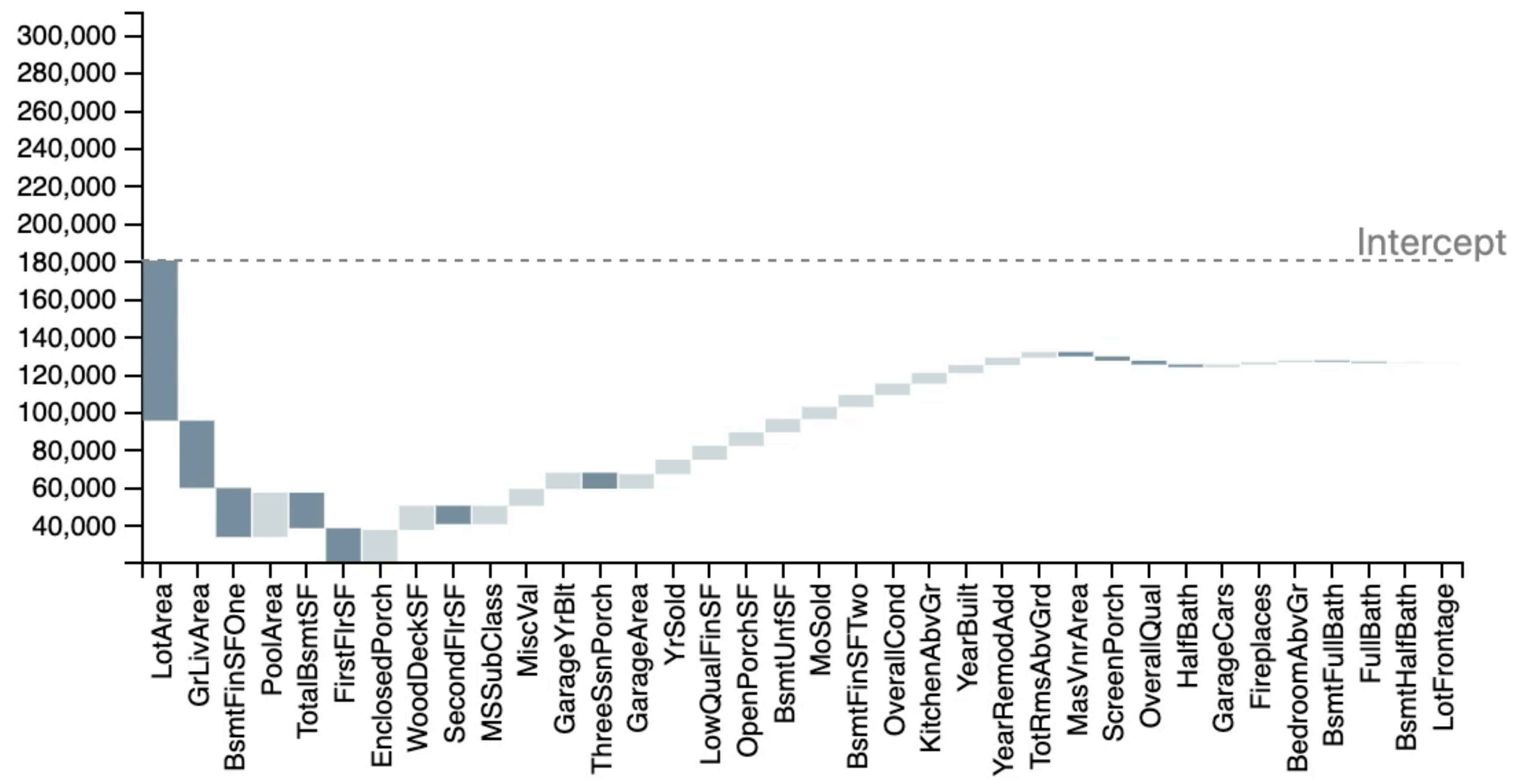
Instance Comparison Summary Settings

Base Instance Summary

Some features have a notable impact on the prediction.

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**



Comparison Summary

Compared Inst

Adjust verbalization **explanation resolution**

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude:

Model Feature Summary

Instance Feature Summary Settings

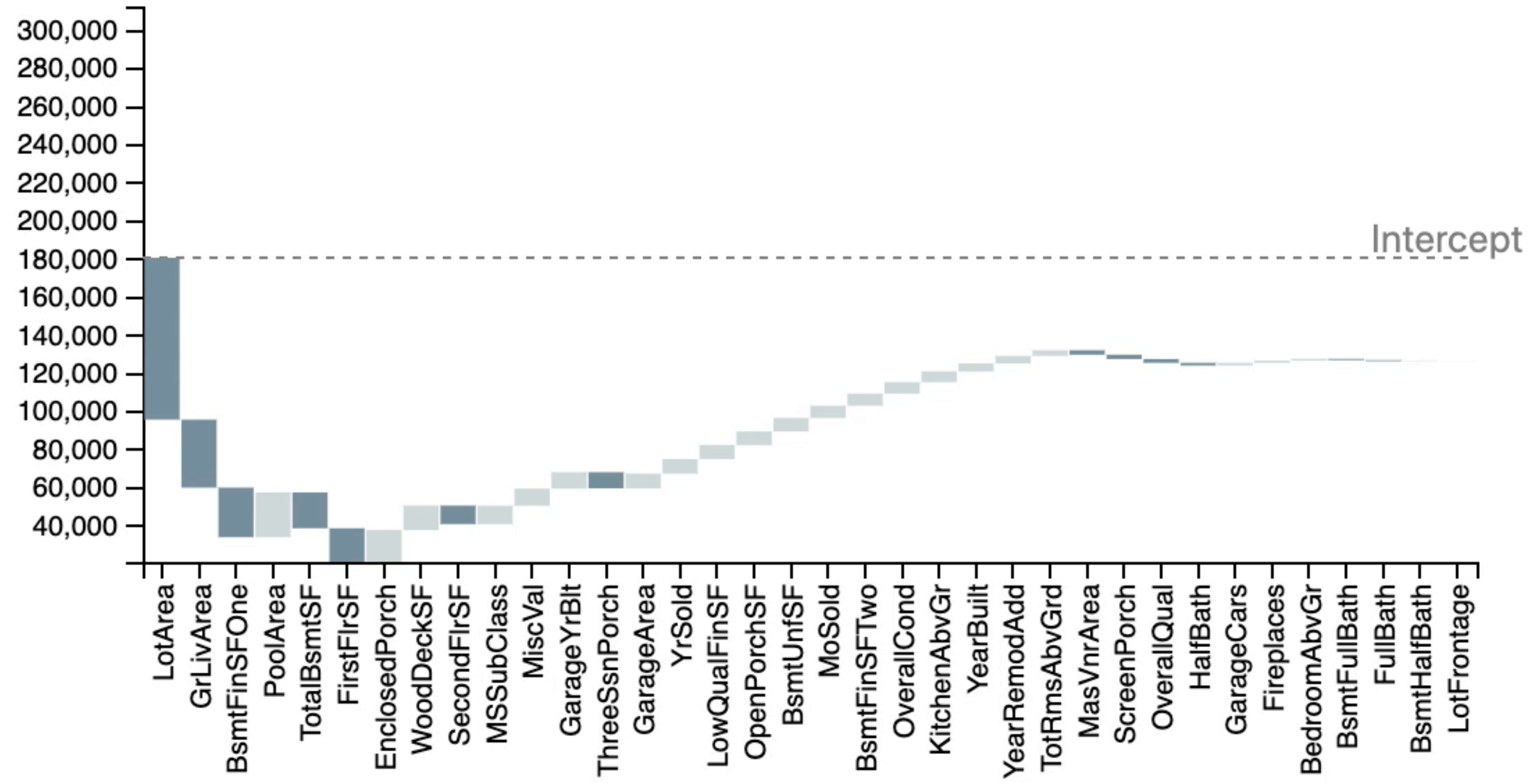
Instance Comparison Summary Settings

Base Instance Summary

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**

Some features have a notable impact on the prediction.



Comparison Summary

Compared Instance Summary

Instance , Actual: , Prediction:

Compared instance:

TeleGam

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude:

Model Feature Summary

Instance Feature Summary Settings

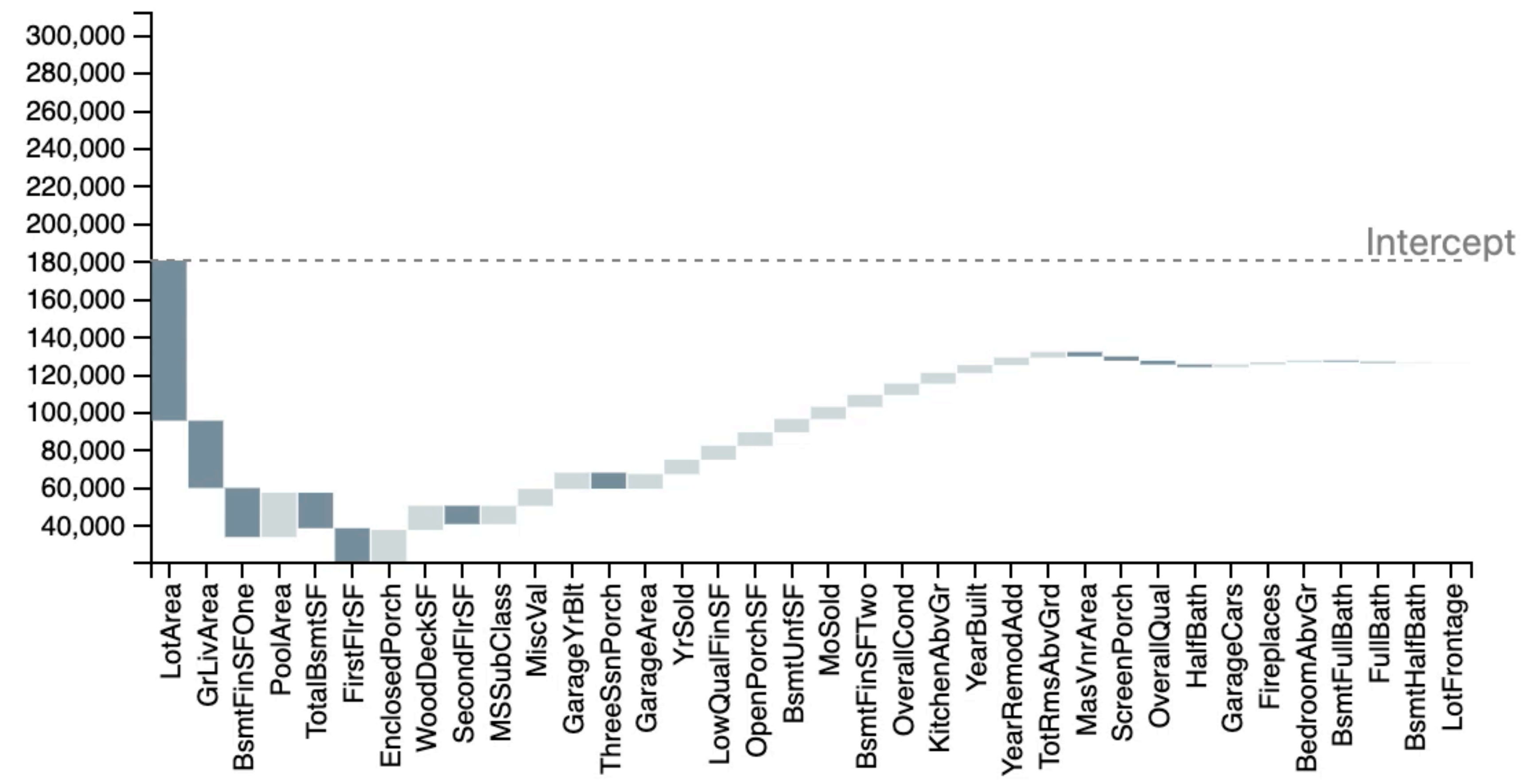
Instance Comparison Summary Settings

Base Instance Summary

Some features have a notable impact on the prediction.

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**



Comparison Summary

Instance , Actual: , Prediction:

Compared instance:

Comparative **verbalization** of two prediction **visualizations**

TELEGAM

Dataset + model: AMES-Housing

Resolution: Brief Detailed

Sort by magnitude: ☑

Model Feature Summary

Instance Feature Summary Settings

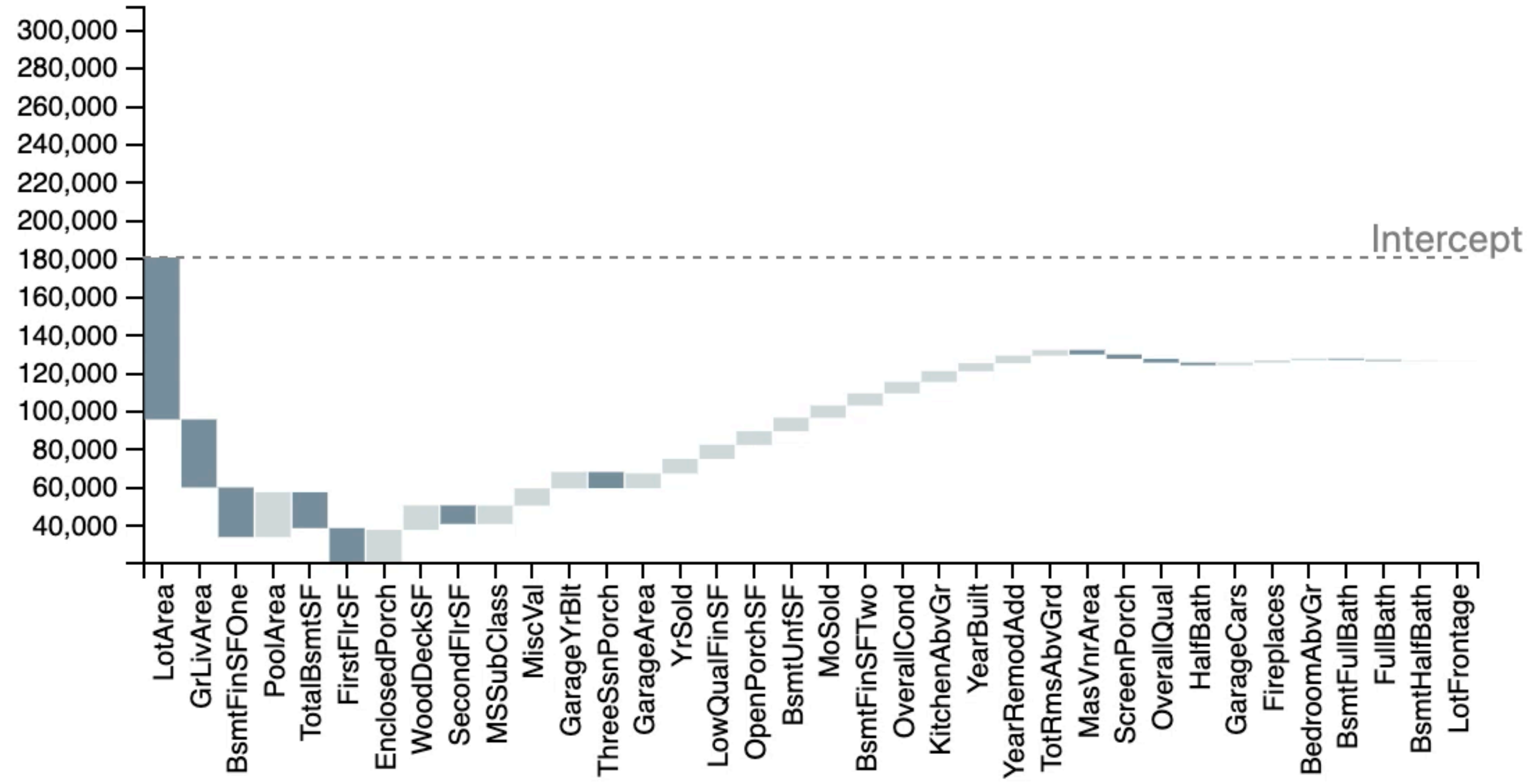
Instance Comparison Summary Settings

Base Instance Summary

Base instance: 7

Instance **7**, Actual: **129900** , Prediction: **126024.98**

Some features have a notable impact on the prediction.



Comparison Summary

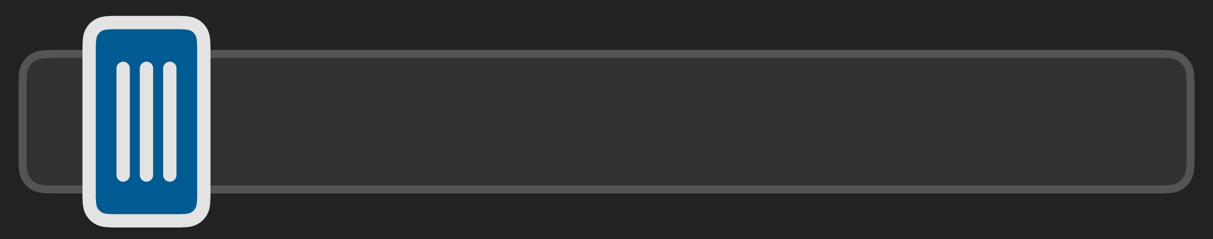
Compared Instance Summary

Instance , Actual: , Prediction:

Compared instance:

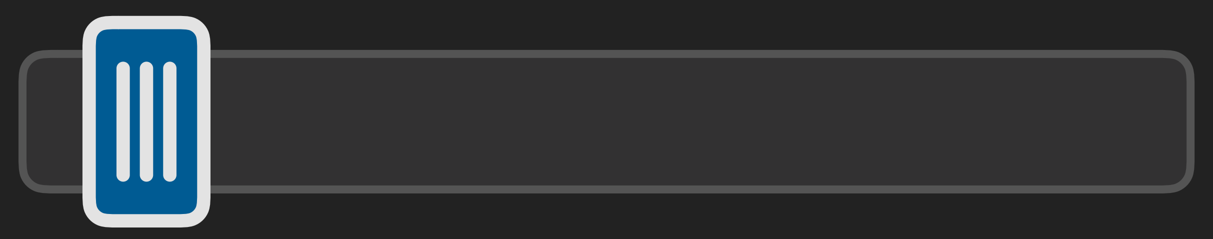
Explanation Resolution

Explanation Resolution

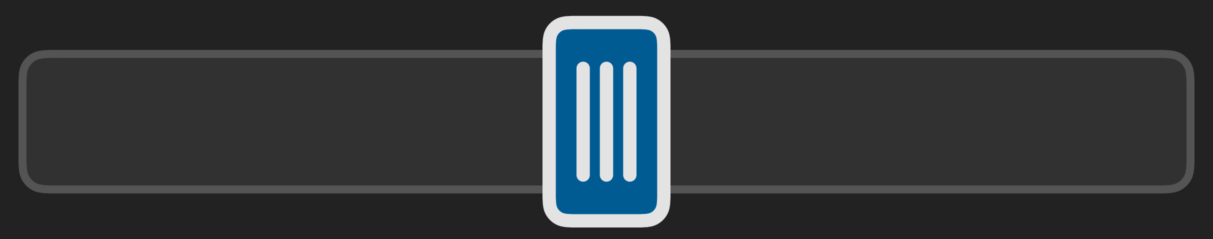


Predictions vary potentially due to ***some features*** *contributing differently from both instances.*

Explanation Resolution

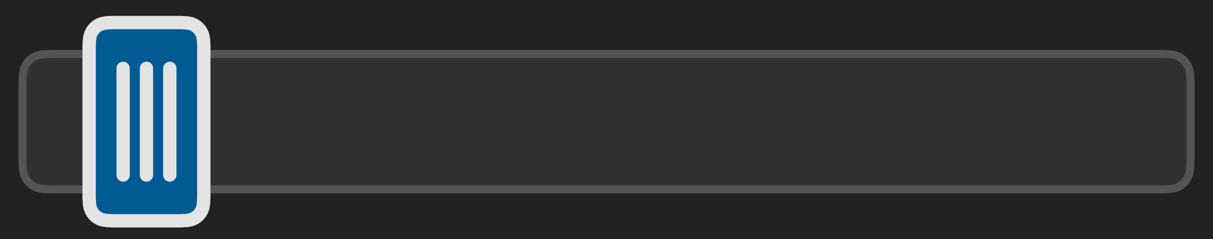


Predictions vary potentially due to ***some features*** *contributing differently from both instances.*

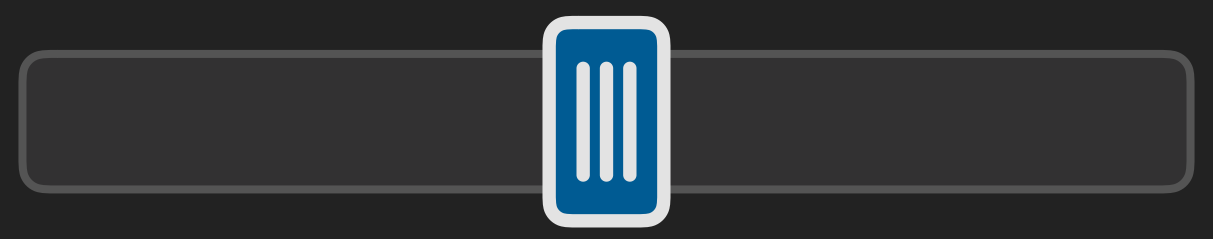


Predictions vary potentially due to ***9 features*** *contributing differently from both instances.*

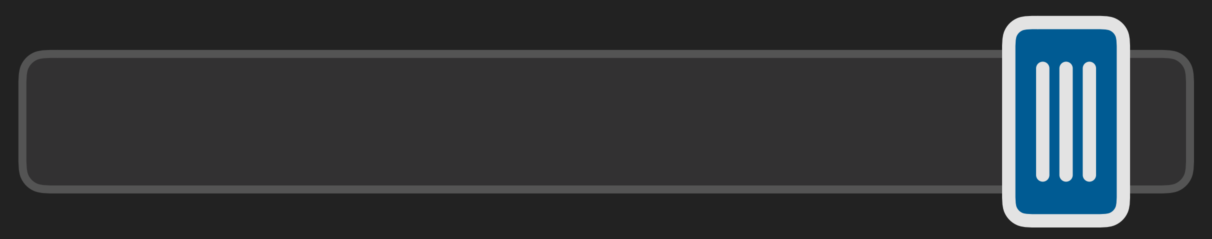
Explanation Resolution



Predictions vary potentially due to ***some features*** *contributing differently from both instances.*



Predictions vary potentially due to ***9 features*** *contributing differently from both instances.*



Predictions ***126,024*** *and* ***312,129*** *vary potentially due to* ***9 features (i.e., 25%)*** *contributing differently from both instances.*

Explanation Resolution

*Predictions vary potentially due to **some features** contributing differently from both instances.*

*Predictions vary potentially due to **9 features** contributing differently from both instances.*

*Predictions **126,024** and **312,129** vary potentially due to **9 features (i.e., 25%)** contributing differently from both instances.*

Brief ⟷ **Detailed**

Verbalization Types

TeleGam	Future Work
Model features	Dataset context
Instance features	Uncertainty
Instance comparison	…

Takeaways

Takeaways

📊 + 📝

Visualization + verbalization are complementary

Combining explanation mediums for the best of both worlds

Takeaways

📊 + 📝

Visualization + verbalization are complementary

Combining explanation mediums for the best of both worlds

💻 + ⚡

Use interaction for generation & presentation

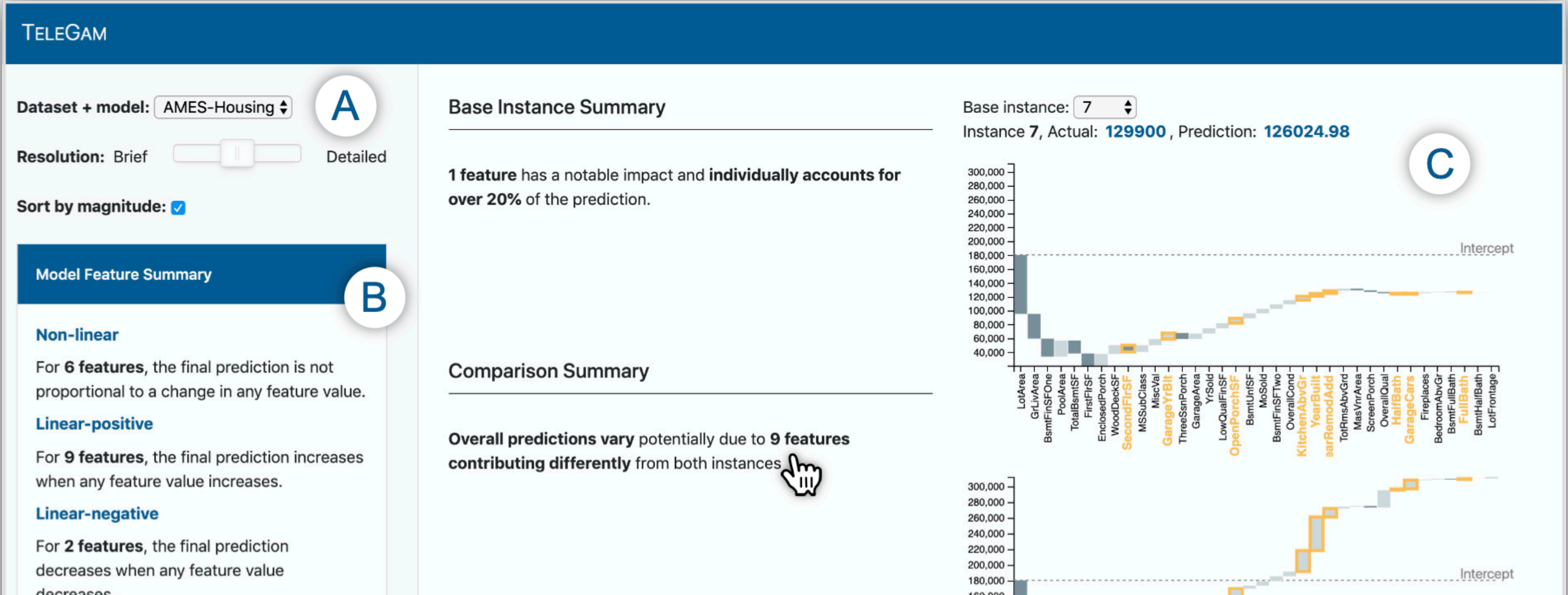
Let users decide resolution, balancing *simplicity* and *completeness*

TeleGam

Combining Visualization and Verbalization for Interpretable Machine Learning

bit.ly/telegam-vis

Demo Paper Video Code Slides



Fred Hohman
@fredhohman
Georgia Tech

Arjun Srinivasan
Georgia Tech

Steven Drucker
Microsoft Research

Thanks!

We thank the GT Vis Lab and the anonymous reviewers for their constructive feedback. Funded by a NASA PhD Fellowship.