Summer Internship Presentation

James Le

August 20th, 2019

DATA JOURNALISM


50%


25%


25%

Agenda

1. Credit Card Insights on Customer Risk
2. Intro to ZAML Monitor
3. The Case for Ensemble Learning

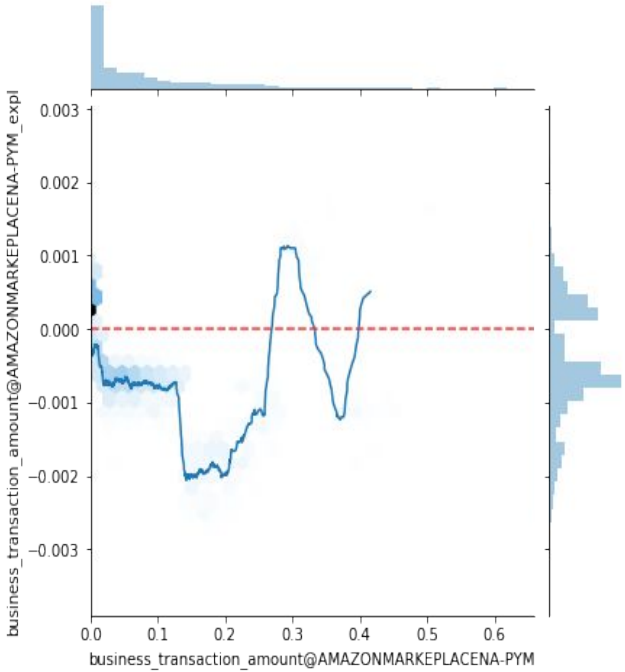
Credit Card Insights on Consumer Risk

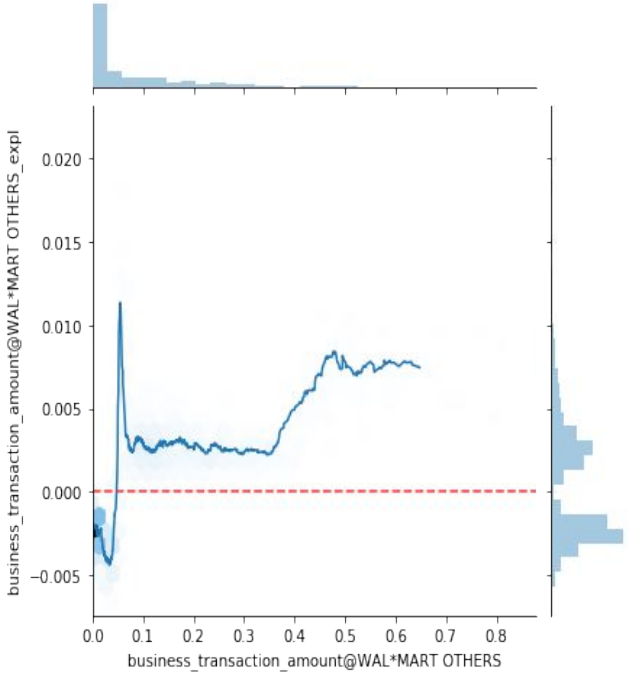
Using ZAML Explain to explore the relationship between purchase behavior and credit risk

Experiment

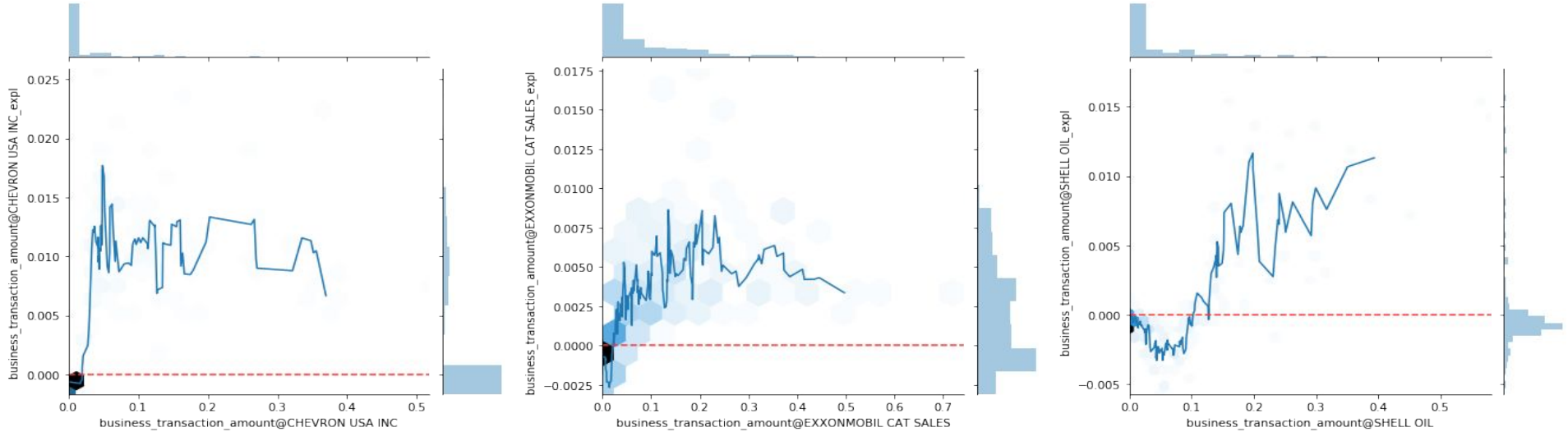
- **Discover Data**:
  - 36,247 customers
  - 701 features
- **Binary Target**:
  - Personal loans
  - 0 - default, 1 - paid
- **Model**:
  - XGBoost
  - 0.61 AUC

Walmart shoppers are riskier than Amazon shoppers





The more you spend at gas stations, the worse your credit risk will be

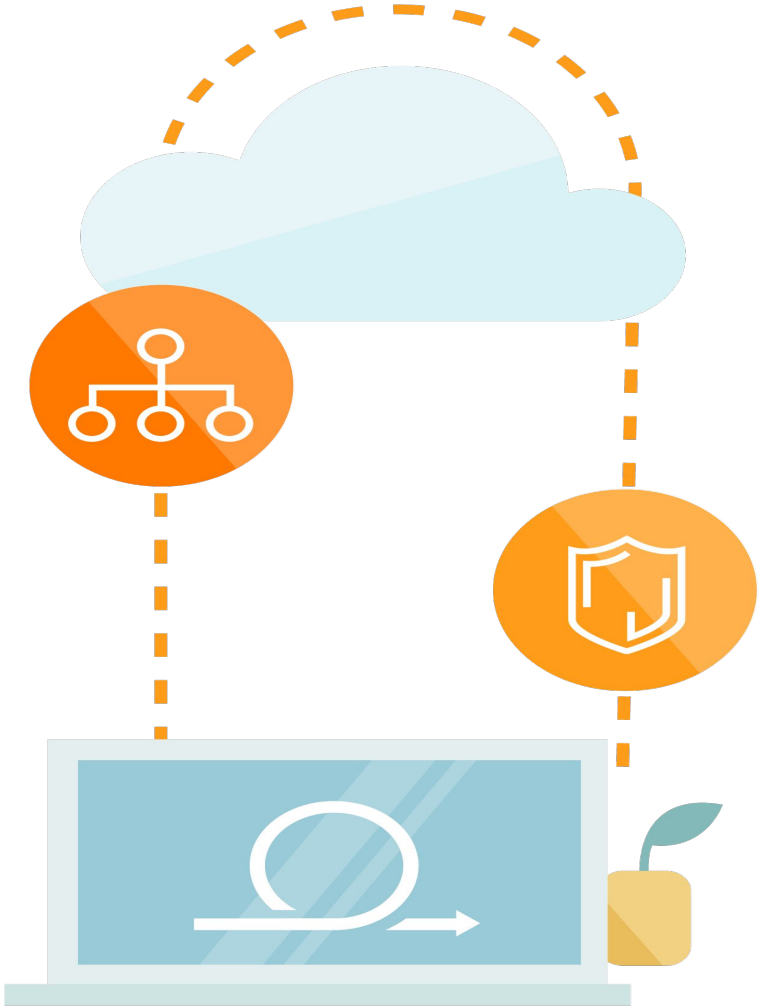


Other Comparisons (> = Higher Risk)

- Transit > Uber > Lyft
- Delta > United > Southwest
- Birchbox > Stitch Fix
- Spotify > Pandora
- GEICO Insurance > State Farm Insurance
- Square > PayPal
- Shopify > Etsy
- Nike > Zappos

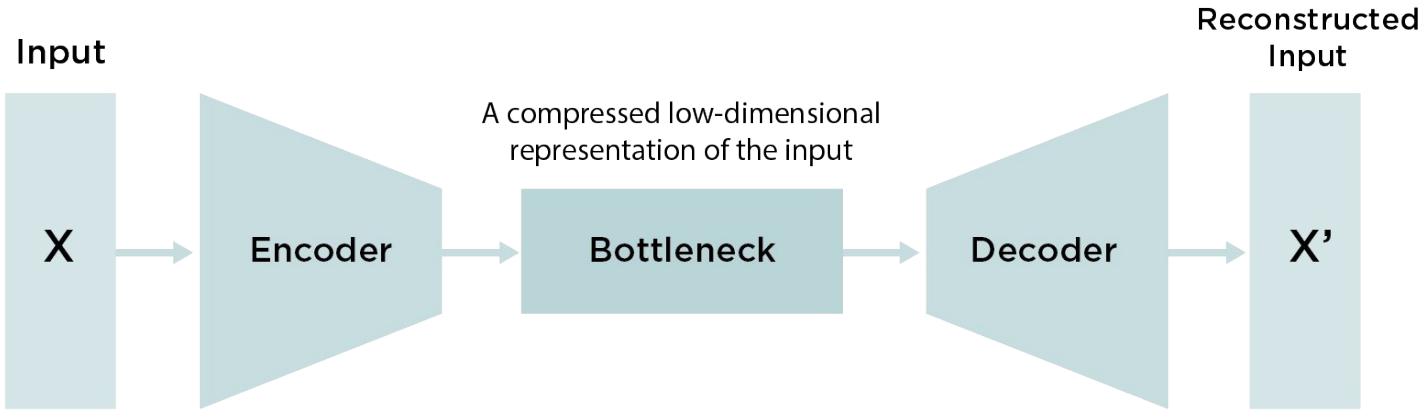
Intro To ZAML Monitor

Using ZAML Monitor to detect changes in model performance



ML Monitoring Approaches

- Concept Drift
- **KS and PSI Tests** look at individual features in isolation
- **Autoencoder (AE)** learn the model behavior
- ZAML Monitor uses a customized AE for the specific ML production model for the specific business problem

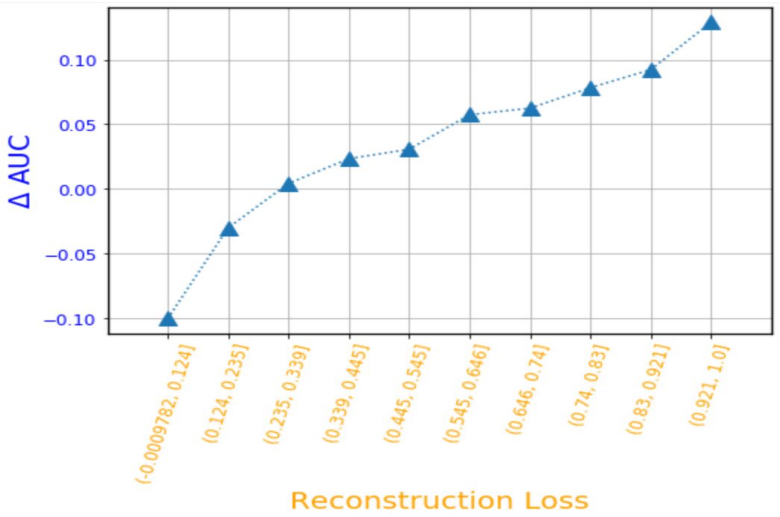


Experiment

- **Akbank Data**:
  - 970,000 customers
  - 3,300 features
- **Binary Target**:
  - Personal loans
  - 0 - delinquent, 1 - paid
- **Model**:
  - Neural Network
  - 0.82 AUC (Validation Set)
- **Monitor:** (Test Set)
  - Standard AE
  - ZAML Monitor's AE

ZAML Monitor is better than common industry practice

Standard Autoencoder



ZAML Monitor's Autoencoder

The Case for Ensemble Learning

More Heads Are Better Than One

Experiment

- **Prestige Data:**
  - >100,000 customers
  - >1,100 features
- **Binary Target:**
  - Auto loans
  - 0 - default, 1 - paid
- **Models**:
  - 4 XGBoost
  - 2 Neural Network
  - Stacked Ensemble

Ensembles Are Better!

Model Type	AUC	KS	Est. Dollars Saved
Ensemble	0.803	0.446	$21M
XGB 1	0.791 (2%)	0.420 (6%)	$18M (14%)
XGB 2	0.791 (2%)	0.428 (4%)	$18M (12%)
XGB 3	0.781 (3%)	0.411 (9%)	$17M (16%)
XGB 4	0.782 (3%)	0.413 (8%)	$17M (16%)
ANN 1	0.750 (7%)	0.376 (19%)	$16M (19%)
ANN 2	0.786 (2%)	0.430 (4%)	$18M (13%)

Challenges with Ensembles

Complexity

Explainability

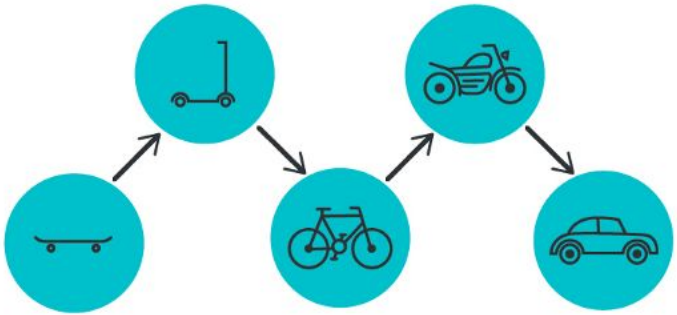
Compliance

Down The Road

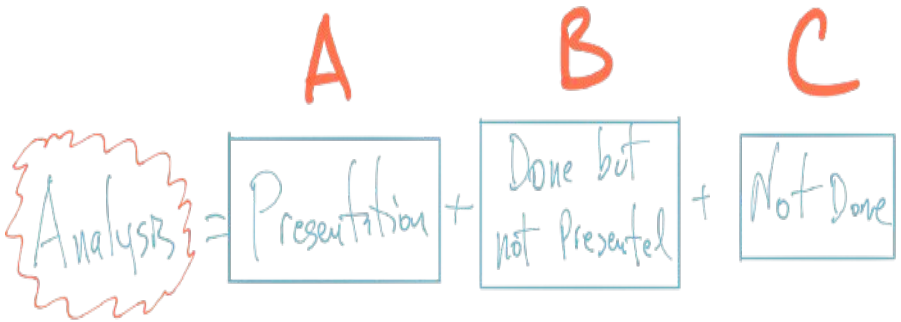
- GIG: Generalized Integrated Gradients (Zest's New Explainability Math)
- ZAML Fair & ZAML Autodoc
- Synthetic Data Generation

Lessons Learned

Minimum Viable Analysis



Shorten The Feedback Loop



Trustworthy Data Analysis

Cognitive Diversity

Diversity of Thought

Lateral Thinking

Impact

Credit Underwriting

Fair Housing

Acknowledgements

Acknowledgements

Acknowledgements

Acknowledgements

Acknowledgements

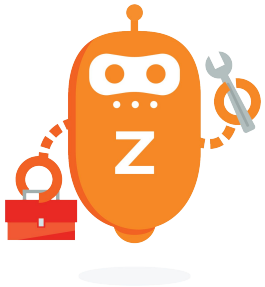
Acknowledgements

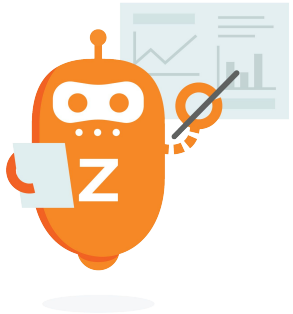
Acknowledgements

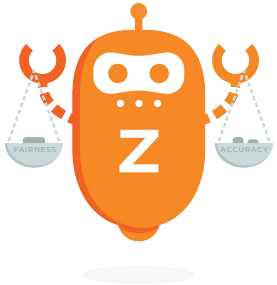
Acknowledgements

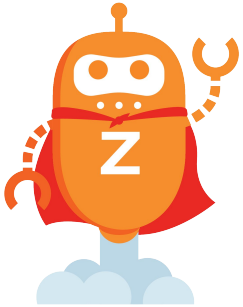
Acknowledgements

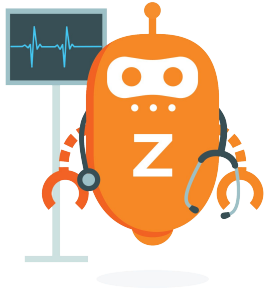
Acknowledgements

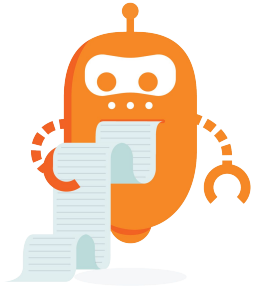





FAIRNESS
ACCURACY







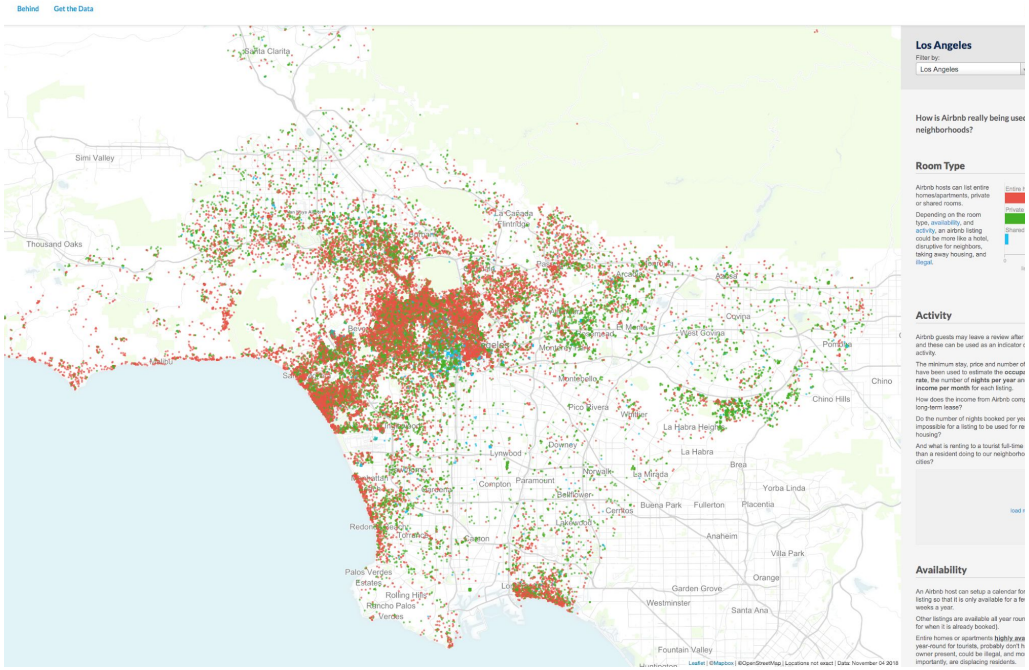
Appendix

What Drives The Price Of An Airbnb Listing?

Using ZAML Explain to reveal the best amenities and the most valuable neighborhoods

Experiment

- **Inside Airbnb Data**:
  - 38,170 listings
  - 564 features
- **Numeric Target**:
  - Price of the listing
- **Model**:
  - Random Forest
  - 0.62 R-Squared



Feature Importance

- Most important factors:
  - Bathrooms
  - Bedrooms
  - Entire Home/Apartment
- Most valuable neighborhoods:
  - Malibu
  - Venice
  - Hollywood Hills
  - Santa Monica

- Most useful amenities:
  - Swimming pool
  - Smoke detector
  - Free parking
  - Indoor fireplace
  - Hot-tub
  - Elevator
  - Gym

Cheaper listings tend to have more amenities and more reviews