Lead Score Model

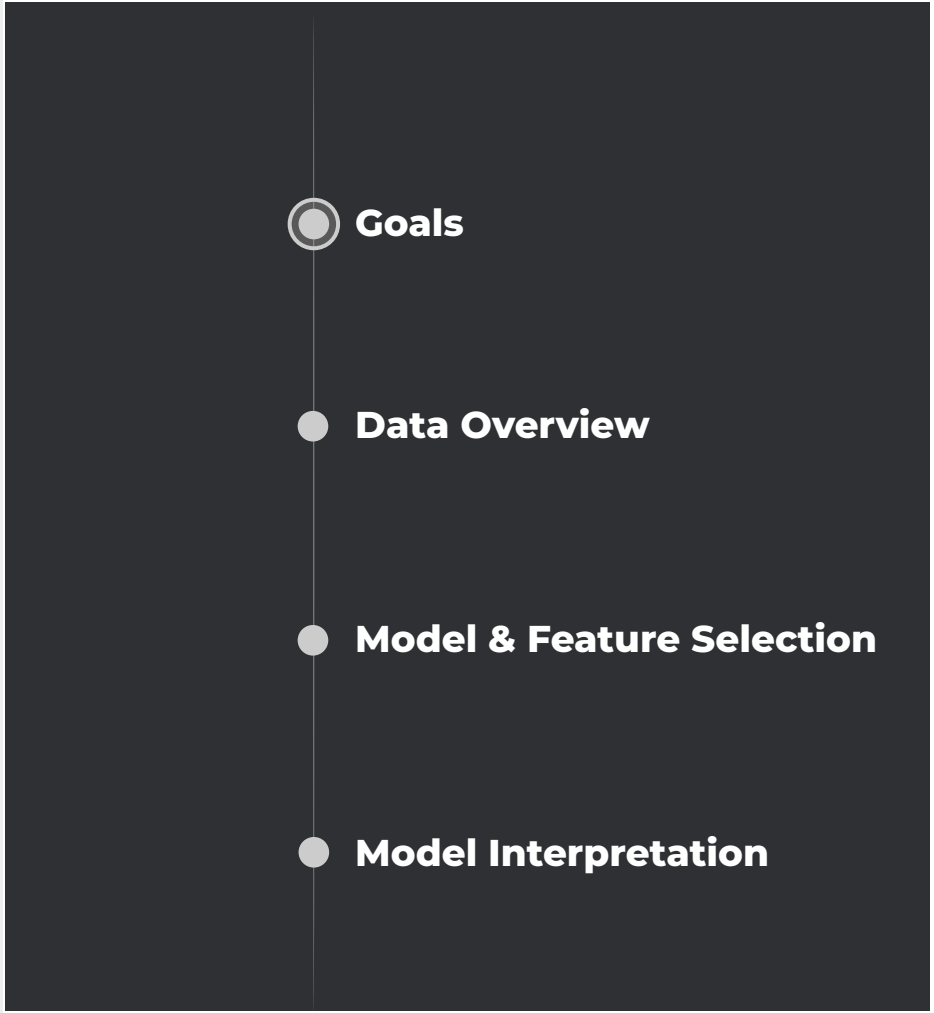
Operations Analytics Intern: Shiyu Ma

Goals:

Predict Conversion Probability of Potential Pros(Leads)

01 Improve Sales Team **Efficiency**

02 **Identify** Characteristics of High Potential Leads



Data Overview

- Goals
- Data Overview
- Model & Feature Selection
- Model Interpretation

Data Overview

3834

Total number of unique leads enriched by sales team (with at least 1 attempt)

62

Features available in dataset

6

Platforms' profile data collected from Internet

7.17%

Conversion Rate

Basic Info

- Industry/Occupation/Category
- State/City/Postal Code/Biz Street
- Phone/Email
- Date Prospected/Enriched/Converted
- Sales Outcome/Number of Attempts

Platforms' Profile

Competitors: Home Advisor/Yelp/Angi/Google

- Profile Link Existence
- Number of Reviews
- Average Rating
- Latest Review Date

Social Media: Facebook/Instagram

- Profile Link Existence
- Number of Followers
- Number of Likes

Stand Alone Website

Basic Info Exploratory Analysis

Conversion Rate by State

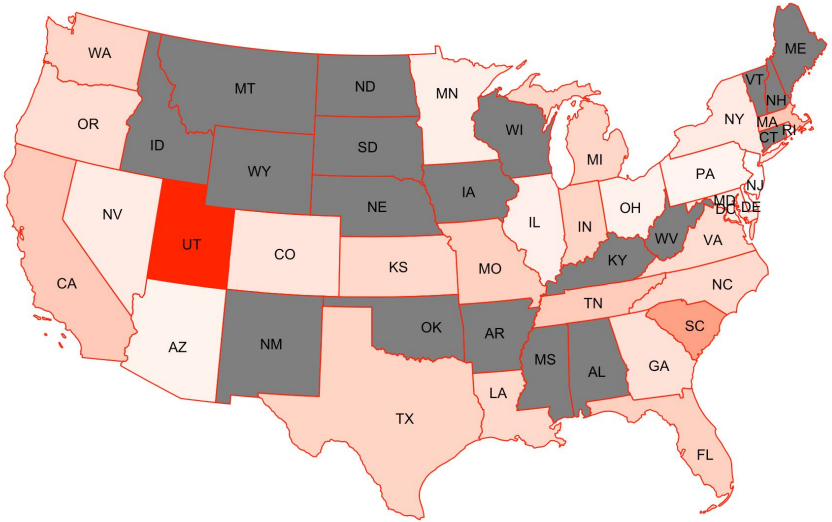
Over 10%: UT, SC, MA, CA, TN, MO, FL

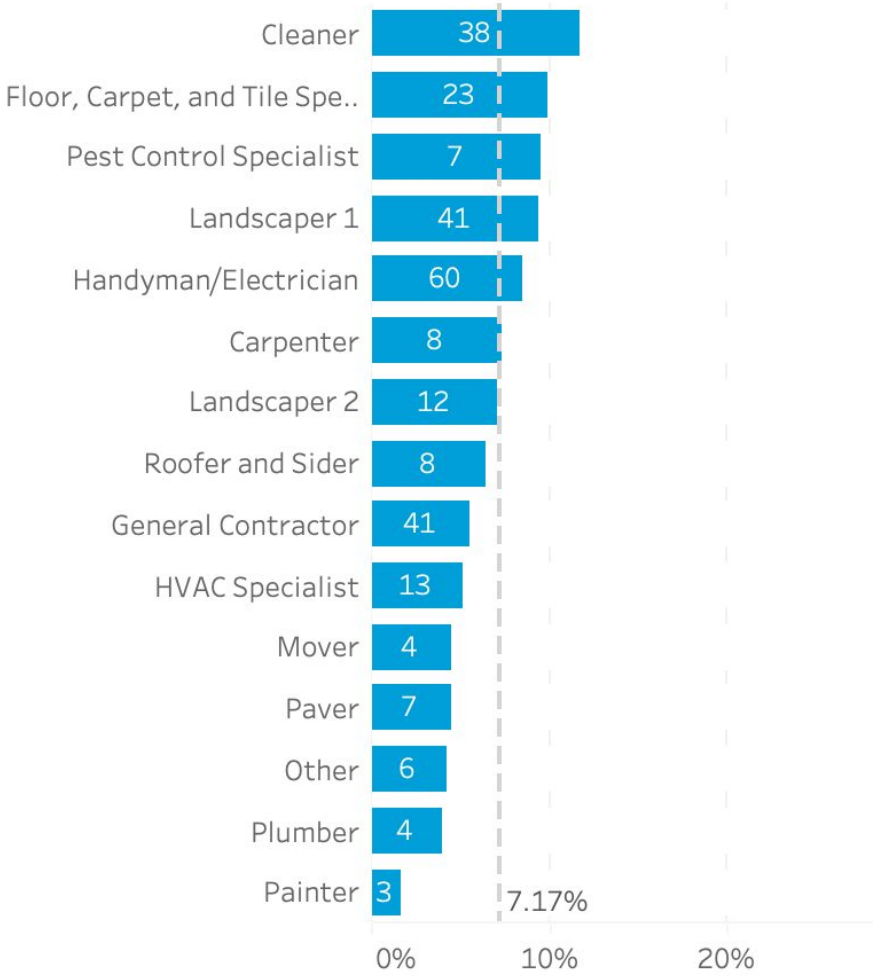
Below 3.5%: RI, NJ, DE, PA, AZ, OH, IL, MN

Conversion Rate by Occupation

Over 9.5%: Cleaner, Floor, Carpet, and Tile Specialist

Below 4%: Plumber, Painter





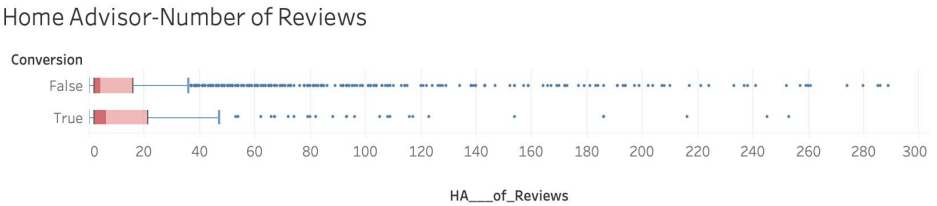
Platforms' Profile Exploratory Analysis

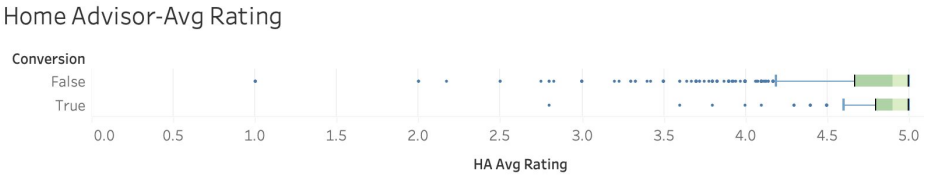
Problems in Features:

- Highly Skewed
- Too Many Outliers

To solve:

- Grouping
- Aggregate between platforms
- Apply model robust to outliers





Data Manipulation Attempts

	Before	After	Why
01	Industry/Occupation/Category	Occupation Group	• **Intrinsic** difference between leads: Service types, Locations • **Grouping** prevent small sample size issue
02	State/City/Postal Code/ Biz Street	**State_G:** State Group	
03	Profile Link Existence	**P_G:** Across Platforms Profile Group	• **Interaction** Between Platforms: Previous Marketing Investment
04	Average Rating & Number of Reviews	Use Wilson interval with continuity correction	• Highlight **Difference** between: Rating 5 (1 review) vs. Rating 4.8 (50 reviews)
05	Latest Review Date	**Diff_Week:** Week Differences from Date Enriched	• Easier format as model input: Convert time into a **numerical** feature
06	Avg. Rating of each platforms	SUM(Avg. Rating)/ Number of Profile Link Existence	**Aggregate** similar features across platforms: • Reduce number of features needed • Less 0, Less skewed data
07	Number of Reviews of each platforms	SUM(Number of Reviews)/ Number of Profile Link Existence	

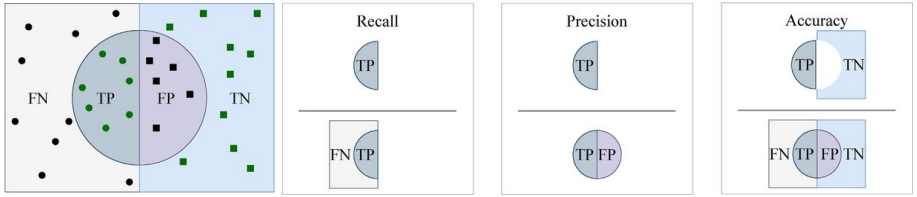
Model & Feature Selection

- Goals
- Data Overview
- Model & Feature Selection
- Model Interpretation

Pre-Model Selection

Imbalance Data Issue:

- Even predicting all samples as FALSE, we still have 93% accuracy=(TP+TN)/All observations



To Solve:

- Use **Recall=TP/Actual P** as evaluation metrics
- Sampling methods help model to **focus** on TRUE samples

Method	Pros & Cons
Stratified Sampling	➢ Built-in for Random Forest ➢ Less training samples per batch
SMOTE Oversampling	➢ No need to delete training samples ➢ Noisy TRUE samples hurt performance
Cost-Sensitive Learning	➢ High Recall=TP/Actual P ➢ Low Precision=TP/Predicted P

Logistic Regression VS. Random Forest

- Use **Random Forest with stratified sampling** as a preliminary model to help with **feature selection** and deal with **messy data**

Method	Pros & Cons
Logistic Regression	➢ Easy to **Interpret** ➢ Fast and Easy to implement
	• Need to ensure no **multicollinearity** exist • Sensitive to **outliers and missing values** • Assume variables linear with logit
Random Forest	➢ Able to handle correlation between features ➢ Robust to outliers and missing values ➢ Able to generate **feature importance**
	• Less interpretable • No available implementation so far

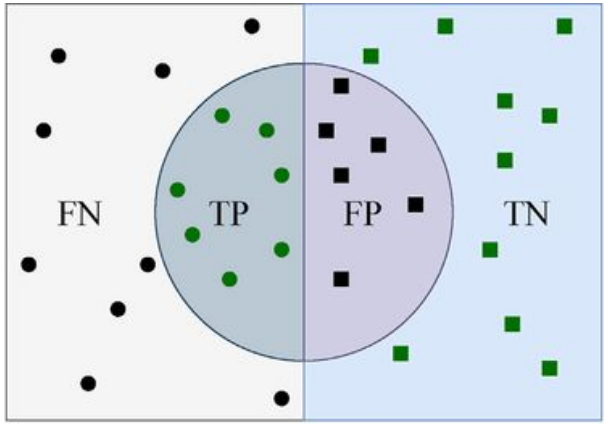
Pre-Model Selection

Imbalance Data Issue:

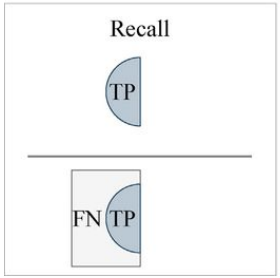
- Even predicting all samples as FALSE, we still have 93% accuracy=(TP+TN)/All observations

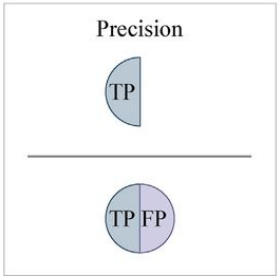
To Solve:

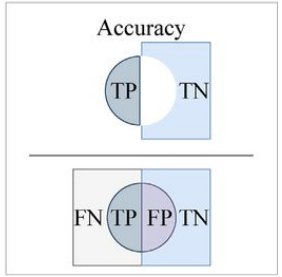
- Use **Recall=TP/Actual P** as evaluation metrics
- Sampling methods help model to **focus** on TRUE samples



Method	Pros & Cons
Stratified Sampling	➢ Built-in for Random Forest ➢ Less training samples per batch
SMOTE Oversampling	➢ No need to delete training samples ➢ Noisy TRUE samples hurt performance
Cost-Sensitive Learning	➢ High Recall=TP/Actual P ➢ Low Precision=TP/Predicted P







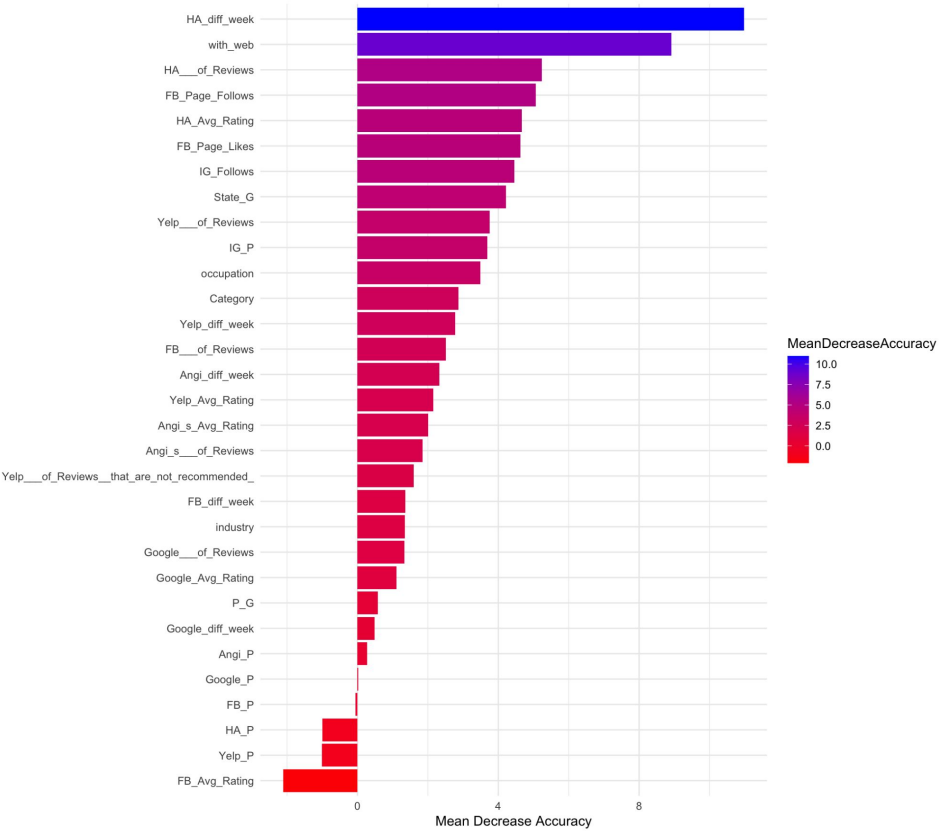
Logistic Regression VS. Random Forest

- Use **Random Forest with stratified sampling** as a preliminary model
- Help with **feature selection** and deal with **messy data**

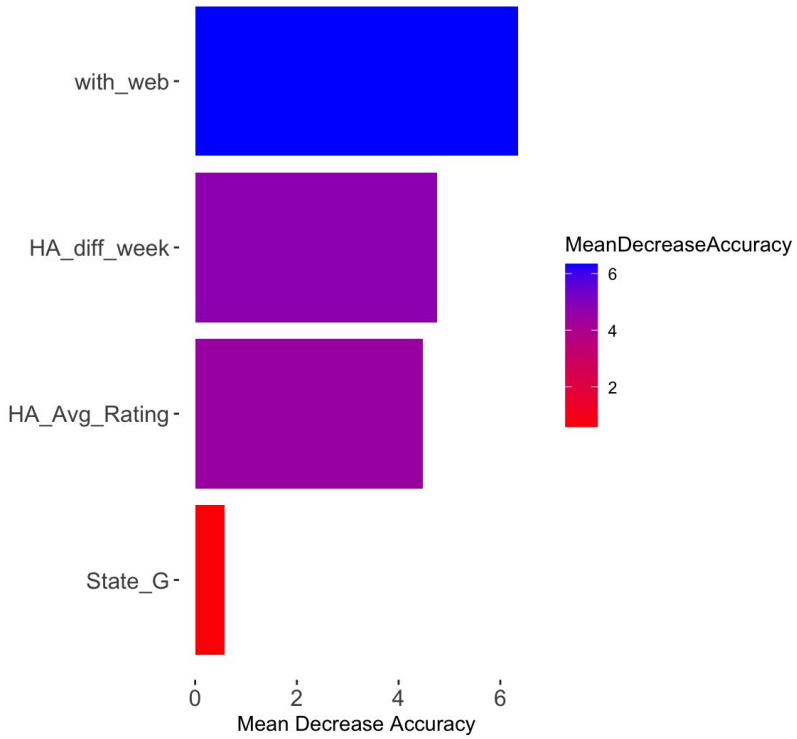
Method	Pros & Cons
Logistic Regression	➢ Easy to **Interpret** ➢ Fast and Easy to implement
	• Need to ensure no **multicollinearity** exist • Sensitive to **outliers and missing values** • Assume variables linear with logit
Random Forest	➢ Able to handle correlation between features ➢ Robust to outliers and missing values ➢ Able to generate **feature importance**
	• Less interpretable • No available implementation so far

Feature Selection

Before: 31 Features
(Median Recall=0.69)



After: 4 Features
(Median Recall=0.72)



Post-Model Selection

Random Forest:

- 4 Features: with_web, HA_diff_week, HA_Avg_Rating, State_Group
- Stratified Sampling
- Higher **Precision**, Lower Recall

Recall	Precision	AUC	ACC
0.72	**0.11**	0.68	**0.58**

	Pred FALSE	Pred TRUE
Actual FALSE	TN=490	**FP=400**
Actual TRUE	**FN=19**	TP=50

Logistic Regression: Chosen

- 4 Features: with_web, HA_diff_week, HA_Avg_Rating, State_Group
- Cost-Sensitive Learning
- Higher **Recall**, Lower Precision
- Regroup HA_Avg_Rating to satisfy **logit linearity** assumption; Remove outliers

Recall	Precision	AUC	ACC
0.93	0.10	**0.72**	0.39

	Pred FALSE	Pred TRUE
Actual FALSE	TN=328	**FP=562**
Actual TRUE	**FN=3**	TP=66

Model Interpretation

- Goals
- Data Overview
- Model & Feature Selection
- Model Interpretation

Flatform Focus in Details

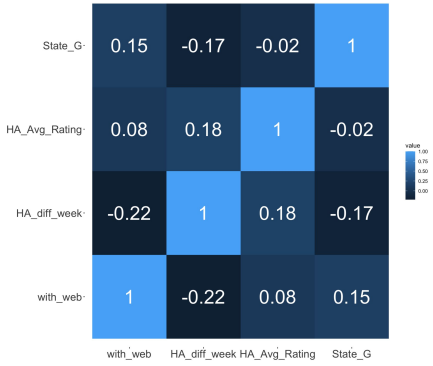
Angi_P	HA_P	Yelp_P	n	c	Conversion Rate
0	0	1	266	33	12.41%
0	1	1	1115	93	8.34%
1	0	1	90	7	7.78%
0	1	0	1698	110	6.48%
1	1	0	164	8	4.88%
1	1	1	494	24	4.86%
1	0	0	7	0	0%

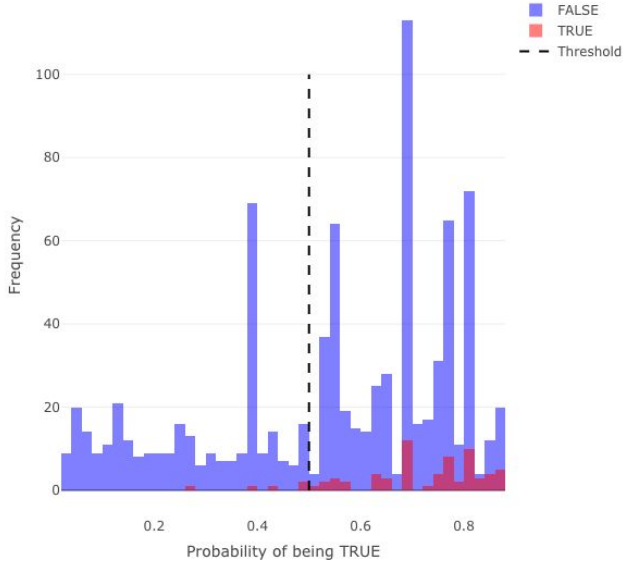
Model Coefficients

Probability of Conversion=

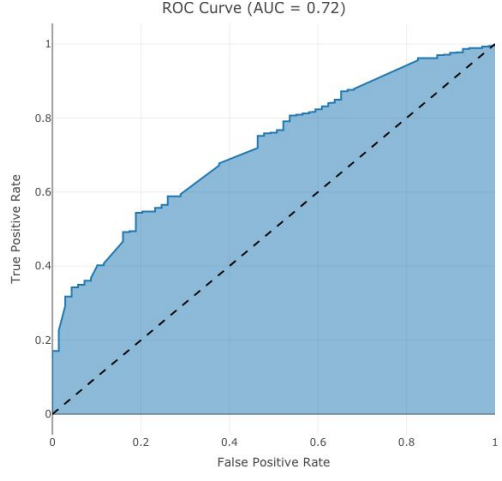
-0.634 + with_web*(0.523) + HA_diff_week*(-0.003)

+ HA_Avg_Rating*(0.375) +State_G*(0.668)





Feature	Possible Values	Coefficient	Interpretation
with_web	FALSE=0, TRUE=1	0.523	For leads with a stand alone website, their conversion probability is multiplied by **1.69** compared to those who does not
HA_diff_week	0 - 1140	-0.003	An increase of 1 week from latest review date means a **0.3% reduction** in the relative conversion probability
HA_Avg_Rating Group	<4.7 = 0, 0 = 1, >=4.7 = 2	0.375	Going up 1 level along 'Low rating, No rating, High rating' means multiplying conversion probability by **1.45**
State_Group	Low = 0, Mid = 1, High = 2	0.668	Going up 1 level along 'Low, Mid, High' Conversion Rate by State group means multiplying conversion probability by **1.95**



Profile of High Potential Leads

Platform Focus
46% are in
Higher Conversion Profile Group
Without Angi + With Yelp

Promising Location
45% located in
Higher Conversion State Group
i.e UT, SC, MA, CA, TN, MO, FL

High Quality
Mean **4.77**
Avg. Rating

Mature
67%
With Stand Alone Website

Active
Median **18 weeks** from
Latest Review Date

Shoutouts 🎉

- **Anne**, for being such a supportive mentor and friend
- **Elliot,** for answering all of my questions and giving me directions
- **Anna**, for leading me to the world of experiments
- **Josh and Blake,** for giving such an interesting project to me and all the constructive feedback
- **The entire Operations Analytics team**, for being so welcoming and helpful
- **All Tackterns**, so nice to talk with you all
- **The University Recruiting team**, for facilitating so many interesting events

Thank you!

Platforms' Profile EDA

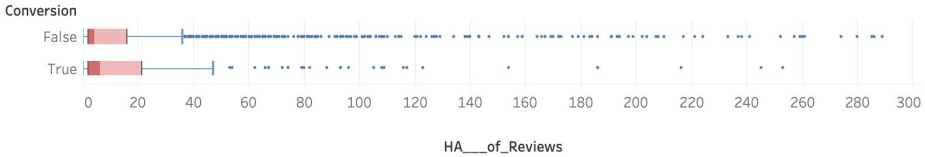
Number of Reviews vs. Avg Ratings

- Highly Skewed
- Too Many Outliers
- No Strong Difference between Conversion
- No Strong Difference between Platforms

Latest Review Year & Stand Alone Website & Across Platforms Profile Link Existence

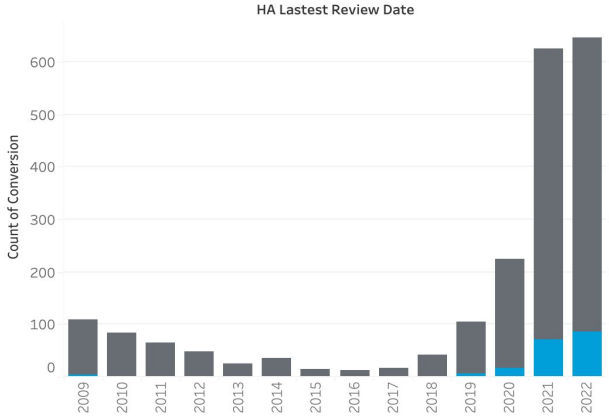
- Promising Features

Home Advisor-Number of Reviews

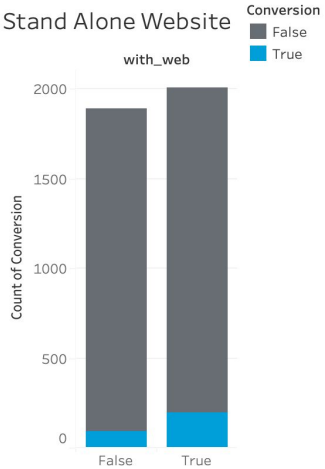


Home Advisor-Avg Rating

HA-Lastest Review Year

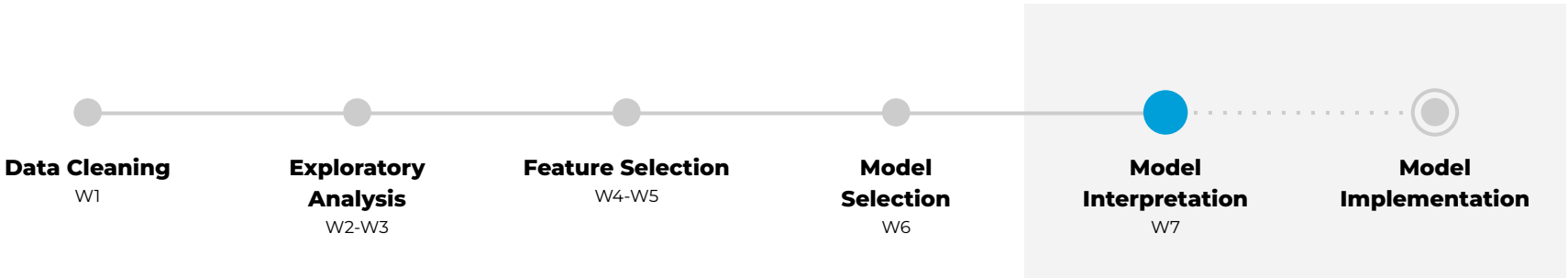


Stand Alone Website



Angi_P	HA_P	Yelp_P	n	c	Conversion Rate
0	0	1	266	33	12.41%
0	1	1	1115	93	8.34%
1	0	1	90	7	7.78%
0	1	0	1698	110	6.48%
1	1	0	164	8	4.88%
1	1	1	494	24	4.86%
1	0	0	7	0	0%

Project timeline



Platforms' Profile EDA

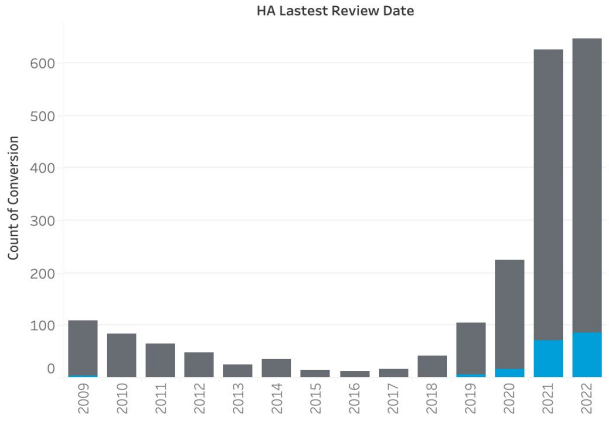
Number of Reviews vs. Avg Ratings

- Highly Skewed
- Too Many Outliers
- No Strong Difference between Conversion
- No Strong Difference between Platforms

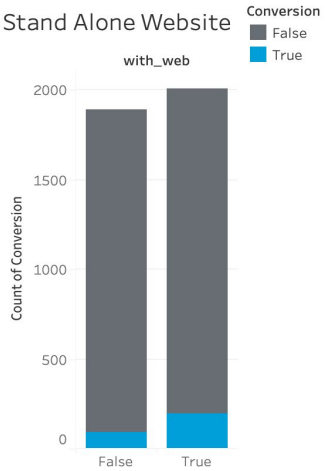
Latest Review Year & Stand Alone Website

- Promising Features

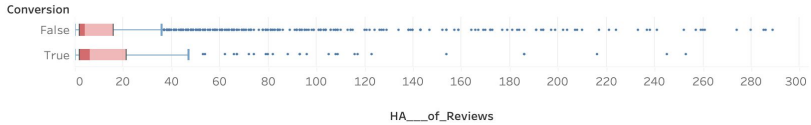
HA-Lastest Review Year



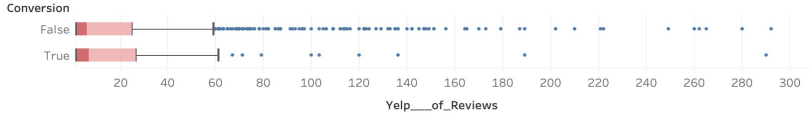
Stand Alone Website



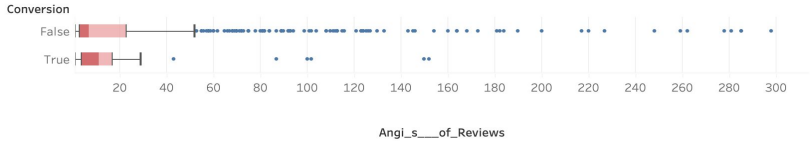
Home Advisor-Number of Reviews



Yelp-Number of Reviews



Angi-Number of Reviews



Home Advisor-Avg Rating

Yelp-Avg Rating

Angi-Avg Rating