Buildertrend Practicum

Yue Ma, Shaheen Nazar, Swetha Vijaya Raju, Aashna Rungta

Buildertrend background

- Buildertrend is a CRM company that provides a software solution for construction companies and for construction material vendors.

- Their business provides an efficient platform for builders to keep track of their jobs, from finalizing a job to purchasing materials to completing construction.

- Buildertrend's clients work across all construction verticals, including residential and commercial construction and they are now the leading project management software for builders, remodelers and contractors.

Contents

- Project scope and objectives
- Benefits to Buildertrend
- Data cleaning
- Task 1: data preprocessing and ML model
- Task 2: data preprocessing and ML model
- Limitations
- Next steps

Project Scope

- The project will mainly focus on developing predictive models to forecast high growth areas and sales prices based on housing demand.
- Analyzing material demand and fair market price based on customers' purchase.
- Apply machine learning and statistical approaches to analyze internal datasets and related external datasets from Census Bureau.

Project Objectives

TASK 1

To analyze housing demand based on location, population and median income and to forecast high growth areas as well as Buildertrend jobs in the future.

TASK 2

To analyze vendor demand based on purchase orders, population, category, region, building permits and building starts

Benefits to Buildertrend

- Forecasting the number of jobs that the company is beneficial so that they can foresee all trends and **adjust their marketing strategies** accordingly, as well as **maintain client expectations** accordingly.
- This would also help them to forecast their business, and to **project financial projections** for their investors in the future
- The **project is beneficial to the close vendor partners** of Buildertrend because the vendors would know what the important features are that could contribute to the improvement of their client-base.
- If Buildertrend could give vendor recommendations to its clients, based on the different features, it could help them **attract more clients** from the construction industry.

Datasets

Internal

- Sample Jobs
- Schedules
- Subs
- Builders
- Purchase Order Items
- Purchase Order Line Items

External

- Population (2019-2021)
- Median income
- Housing Starts
- Building Permits

SampleJobs

This dataset captures data entered by builders into the Buildertrend platform and contains information on each of the jobs that they had or working on, includes location, date opened, project type, and approximate starting price.

```
df_BT_SampleJobs.head()
```

	builderID	jobID	jobName	dateOpened	projectType	street	zip	city	State	startingPrice
0	2	4891421	Hall Residence	2019-01-08T17:17:24.997Z	New Home - Custom	12903 Traceview Loup	51503	Council Bluffs	IA	481140.0
1	2	5052030	Swanson/Faraci	2019-02-07T16:11:18.527Z	Commercial - Renovation/Tenant Improvements	12917 Heavenly Dr	68154	Omaha	NE	780000.0
2	2	5128567	Willey Residence	2019-02-25T17:50:41.287Z	New Home - Custom	TBD	68023	Fort Calhoun	NE	461419.0
3	2	6124332	Schutt Residence	2019-09-04T15:28:07.623Z	New Home - Custom	6220 State St	68152	Omaha	NE	NaN
4	2	6506201	Wickham Residence	2019-11-06T17:57:43.330Z	New Home - Custom	11303 N 126th St	68142	Omaha	NE	550050.0

Schedules

This dataset contains more in-depth information on each job that builders have, contains details of each job timeline, and includes 15 columns.

	builderID	jobsiteID	scheduleID	title	enteredDuration	actualDuration	workDays	isMarkedComplete	dateEntered	startDateTime	endDateTime
0	2	4891421	62391307	Ordered Steel Beams	1	1	62	False	2019-01-08 18:10:57.923	2019-01-08 06:00:00.0	2019-01-09 05:59:00.0
1	2	4891421	62391369	Deliver Beams	1	1	62	False	2019-01-08 18:11:42.627	2019-01-14 06:00:00.0	2019-01-15 05:59:00.0
2	2	4891421	62391458	Framing	33	45	62	False	2019-01-08 18:12:20.487	2019-01-14 06:00:00.0	2019-02-28 05:59:00.0

Subs

The dataset has the vendor names which builders contact for their material purchase.

df_subs

	builderID	subID	subCompanyName
0	2	15281	Builders Supply
1	2	15283	The Frazier Co. Nate
2	2	15349	Rock Solid
3	2	15352	Century Electric
4	2	15353	Carlisle Insulation

Pattern Matching

- Identify different patterns using regular expressions.
- Create a bag of all pattern results.
- Replace them with the right vendor name.

Subs

```
df_partner_vendors.subCompanyName.value_counts()
```

Home Depot	5026
Lowe's	1949
84 Lumber	1882
Ferguson Enterprises	1046
Lansing Building Products	663
HD Supply	535
Builder's First Source	221
Beacon Building Products	208
McCoy's Building Supply	151
Consolidated Electrical Distributors	51
SRS Distribution	7
Graybar Electric Co.	4
Hajoca Corp.	4
MRC Global	2

```
df_parnered_companies.subCompanyName.value_counts()
```

Ferguson Enterprises	8760
Home Depot	8519
Lowe's	7057
Builder's First Source	3743
84 Lumber	2971
HD Supply	1392
Lansing Building Products	812
McCoy's Building Supply	667
Hajoca Corp.	515
Consolidated Electrical Distributors	453
Winsupply Inc.	320
Graybar Electric Co.	302
Beacon Building Products	290
Rexel Holdings	250
WESCO International Inc.	110
MRC Global	93
Anixter International	85
Emco Corp.	56
SRS Distribution	29
US LBM	23

Builders

The dataset has all the builders' details on the Buildertrend platform

	builderid	builderName	siteDomain	primaryPhone	primaryEmail	street	city	state	postalCode	country
0	77497	Frontier Concrete & Masonry LLC	NaN	912-508-4900	miranda@frontierconcrete.org	2209 Rowland Ave	Thunderbolt	GA	31404	US
1	77121	Down Under Construction	NaN	(801) 936-2400	katie.wheeler@downunderconut.com	590 900 North	North Salt Lake	UT	84054	US
2	76678	Cuates Construction	NaN	9567359722	cuatesconstruction@yahoo.com	PO 822	rio grande city	TX	78582	US
3	77452	RI Construction Consulting Company Inc	NaN	2093288613	roxsan.perez@gmail.com	18373 Exeter Place	Lathrop	CA	95330	US

Purchase Order

The dataset stores user input order brief information and the timestamp when the user entered the record.

```
df_orders.head()
```

	builderID	jobID	purchaseOrderID	purchaseOrderTitle	subID	dateAdded
0	53	6682579	20286840	Plumb rough	5395063	2020-03-24T01:25:15.640Z
1	53	6682579	20286852	Post tension cables	5395176	2020-03-24T01:26:38.450Z
2	53	6682579	20286867	Pad and footing	5395318	2020-03-24T01:28:13.483Z
3	53	6682579	20286877	Concrete footing	5395034	2020-03-24T01:29:33.967Z
4	53	6682579	20286883	Slab and fill	5395318	2020-03-24T01:30:47.390Z

```
df_orders.shape
```

```
(11114957, 6)
```

Purchase Order Line Items

This dataset contains order items for each order.

```
df_orderline.head()
```

	builderID	jobID	purchaseOrderID	PurchaseOrderLineItemId	costCodeTitle	lineItemAmount
0	49626	8973189	25608497	40383846	4502-Draw #2	0.0
1	49626	8973189	25608503	40383852	4120-Rockwall	0.0
2	49626	8973189	25608517	40383866	4401-Rough	0.0
3	49626	8973189	25608520	40383869	4011-HVAC	0.0
4	49626	8973189	25608523	40383872	4005-Roofing	0.0

```
df_orderline.shape
```

```
(17873291, 6)
```

External Datasets

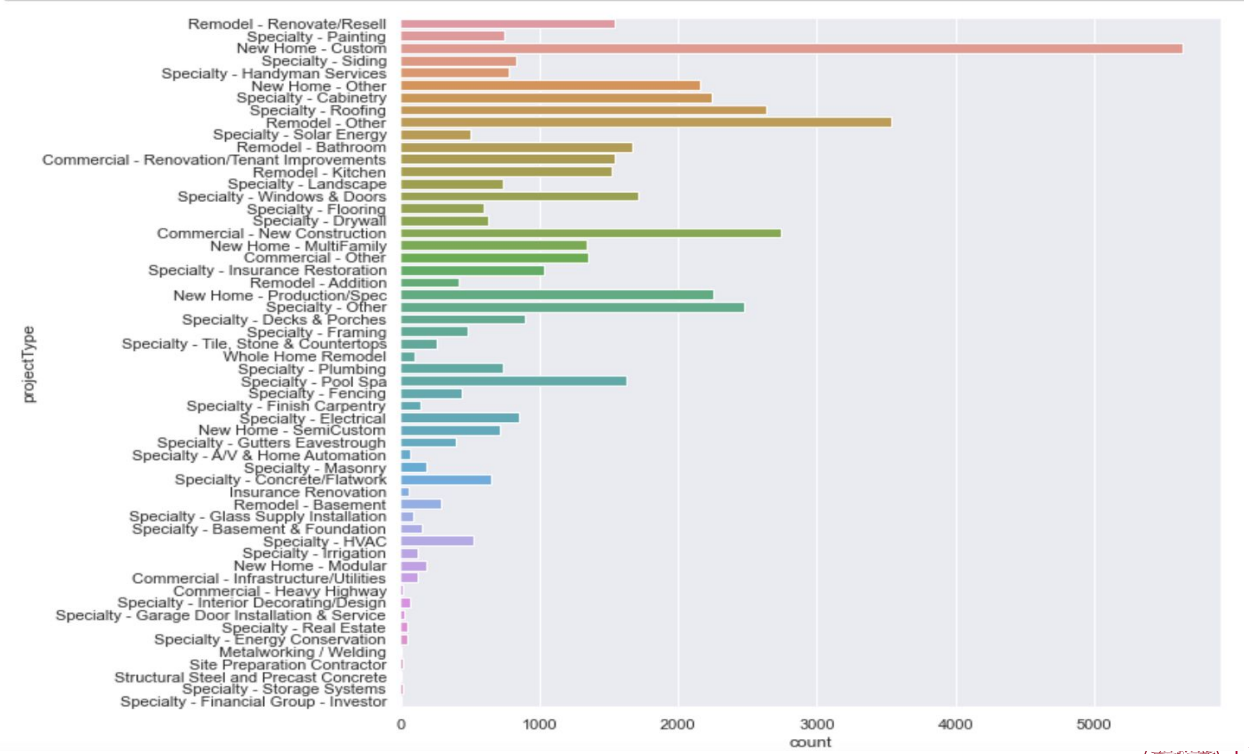
- All taken from the US Census Bureau:
  - Population per region per year
  - Median income in every household type per region
  - Number of Housing starts per region per month
  - Building Permits per state per month

Task 1

Predict number of Buildertrend jobs per region per month

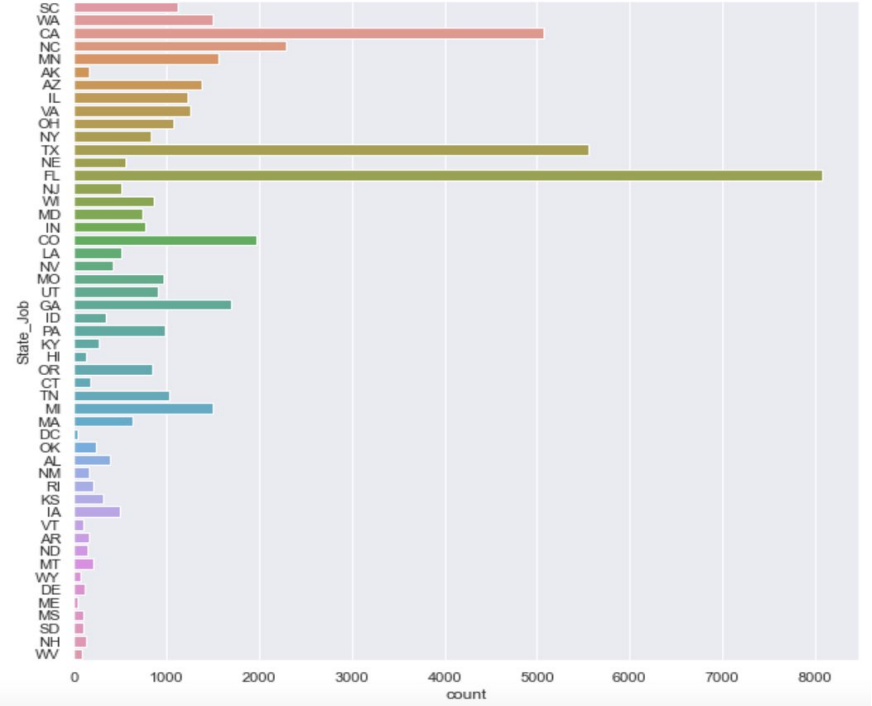
Exploratory Data Analysis

Most popular project types at Buildertrend

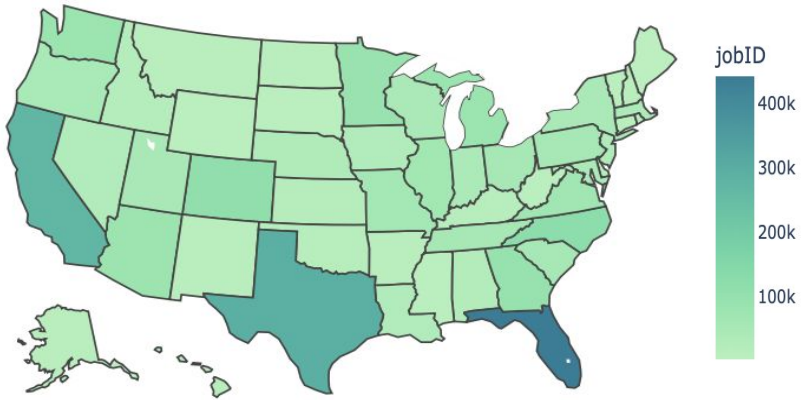


Exploratory Data Analysis

States where Buildertrend has the most jobs



Number of jobs per state



Data Preprocessing

- Data cleaning:
  - Number of states in the original sampleJobs dataset: 1481
  - Number of states after cleaning: 51
- We only looked at the New Home constructions:
  - Because that's the only data we had from the Census Bureau
  - Because it's the most popular category by far at Buildertrend
- States were all divided into 4 regions:
  - West
  - Midwest
  - South
  - Northeast
- Household Type:
  - Families
  - Households
  - Married-couple families
  - Nonfamily households

Final Model

- Variables of the model:
  - Month
  - Total number of housing starts in the region
  - Mean income within that household type within that region
  - Household type
  - Population of the region
- Using these, we are trying to forecast number of jobs that Buildertrend gets in the New Home Construction Category in any region.

ML Model

- AutoGluon - It's a multimodal that takes all the features, identifies the problem type and runs it through 7-8 different models to give you the best/most accurate result.
- Results:
  - Each of our variables that we used are all very important in determining the number of jobs that Buildertrend will get.
  - **We can use this model to forecast number of jobs per region with 99% accuracy**
  - Gives us info on how the company will do in the new home category, so we can use this for a few things:
    - Financial projections
    - Marketing - where/who to market to next
    - Informing vendors about material demand in the future

Limitations & Next Steps

Limitations:

- The model only works if we're forecasting for a time period where nothing major happens
- Would like to test it out with more variables - age distribution, changes in the economy

Next steps:

- Try to run the same one but with different project types - maybe with commercial construction
- Try to find more granular external data, especially with states/regions
- Test out other variables in the model to see if we can find more things that improve the accuracy
- Get more data over time to see how Buildertrend's numbers react to different 'abnormal' situations

Task 2

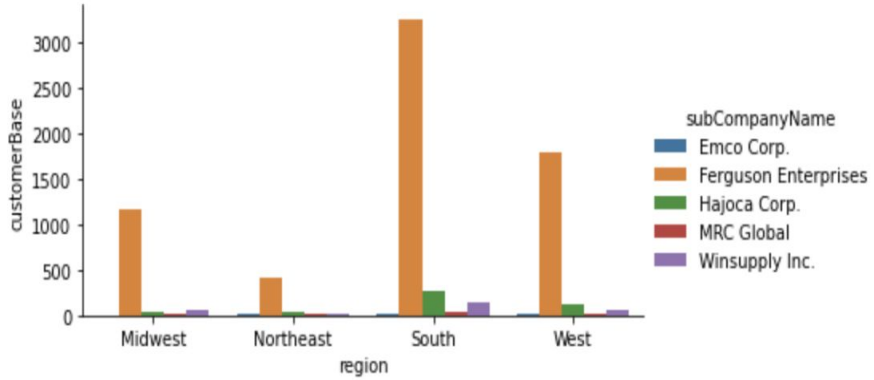
Predict the Customer Count per month statewise for each vendor

Exploratory Data Analysis

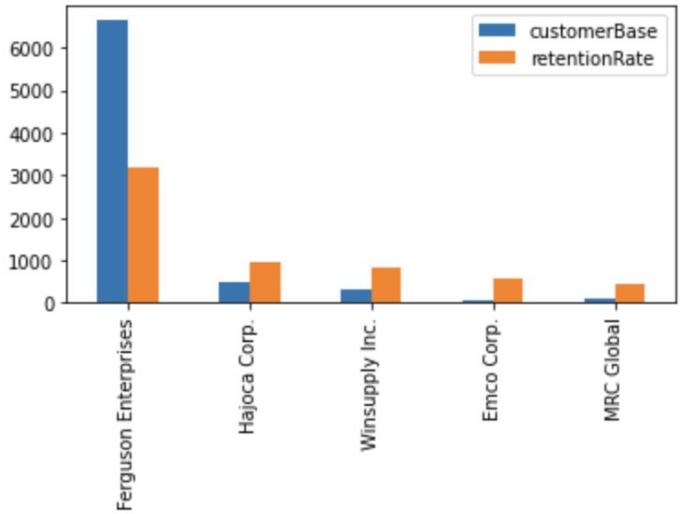
- Calculated the **Customer Base** for each vendor from the unique Builders statewise.
- Customer Index for each vendor state wise since Customer Base is proportional to population. This serves as a better metric to understand the customer spread taking into account statewise population.
    - **Customer Index = total number of customer for each vendor / total population**
- Calculated a retention rate metric for each vendor based on total number of orders and customer base.
    - **Retention Rate = Total number of orders for a vendor / customer base of vendor**
- Found the most **popular vendor** region wise for each category.

Popular Vendor - Plumbing

Customer Base

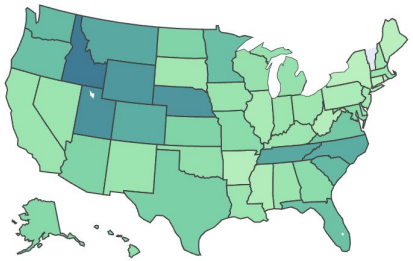


Retention Rate

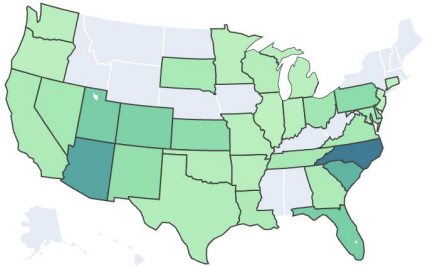


Customer Index - Plumbing

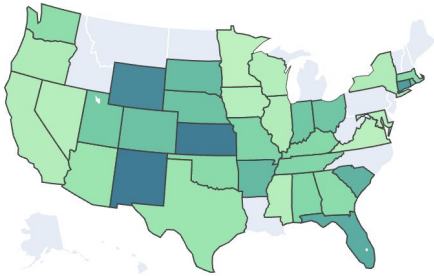
Ferguson Enterprises



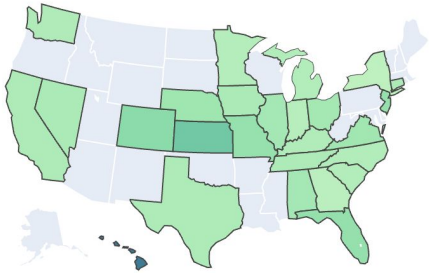
Hajoca Corp.



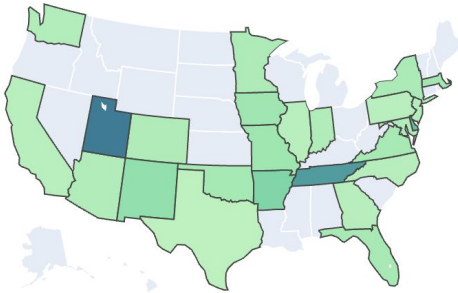
Winsupply Inc.

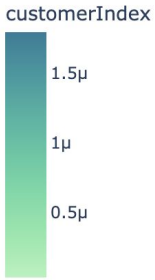


MRC Global



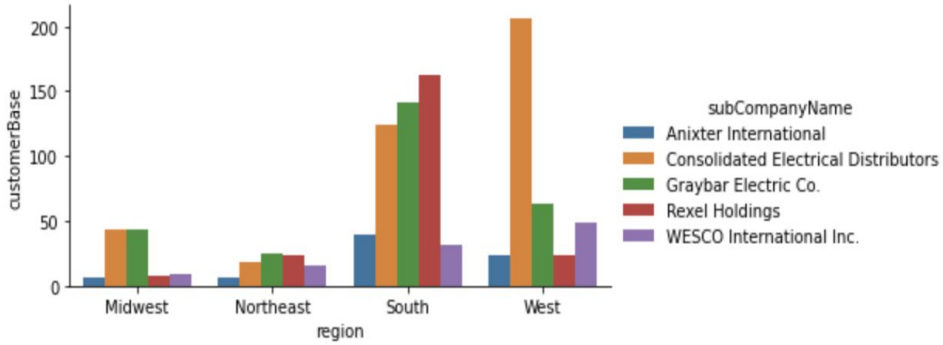
Emco Corp.



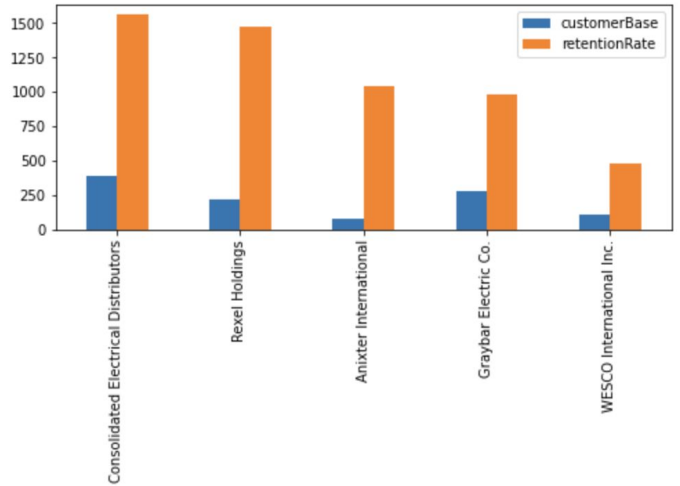


Popular Vendor - Electrical

Customer Base

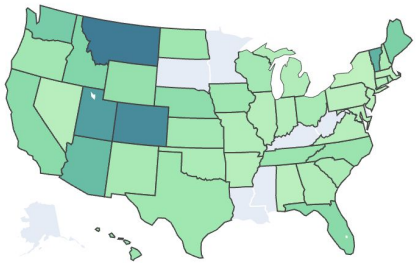


Retention Rate

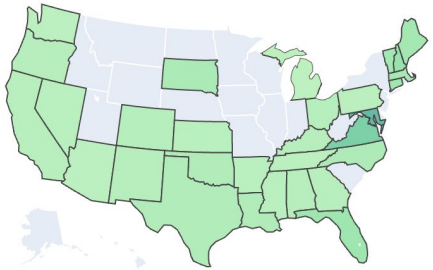


Customer Index - Electrical

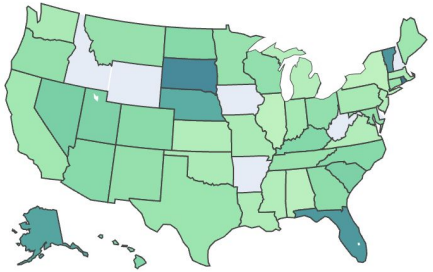
Consolidated Electrical Distributors



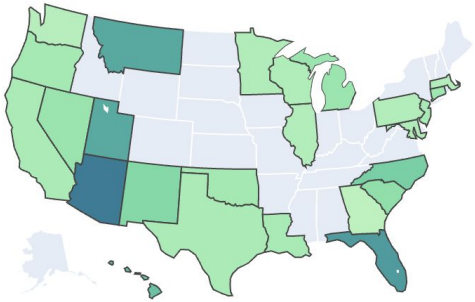
Rexel Holdings



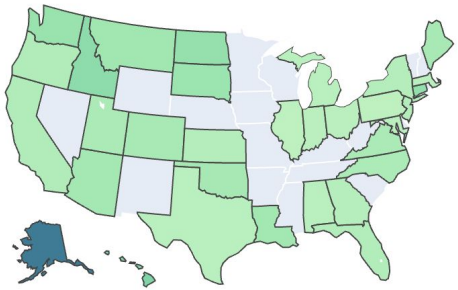
Graybar Electric Co.

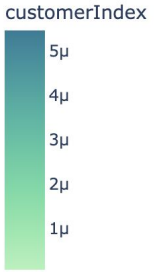


Anixter International



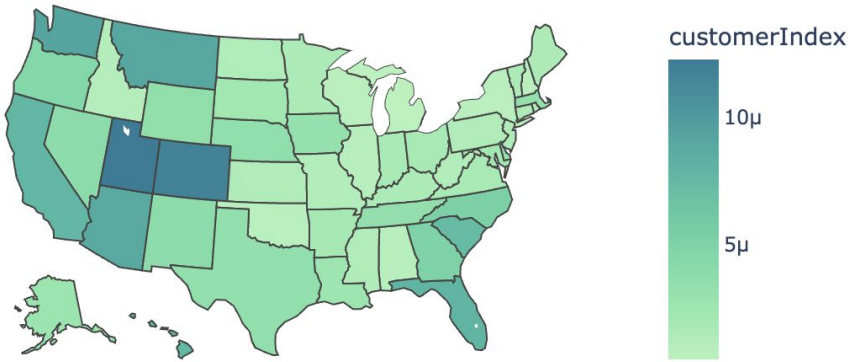
Wesco International Inc.



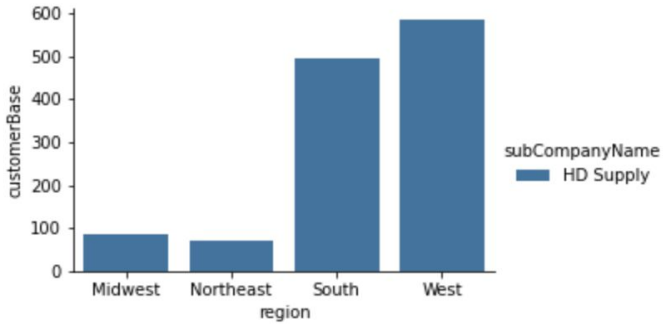


Customer Index - Electrical & Plumbing

HD Supply

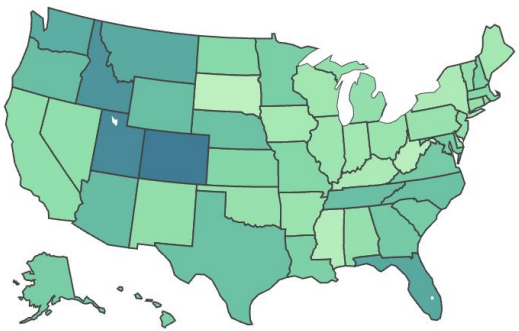


Popular Vendor

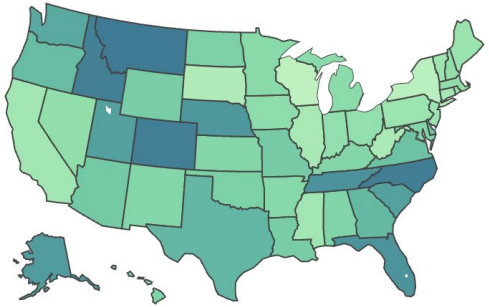


Customer Index - Home Center

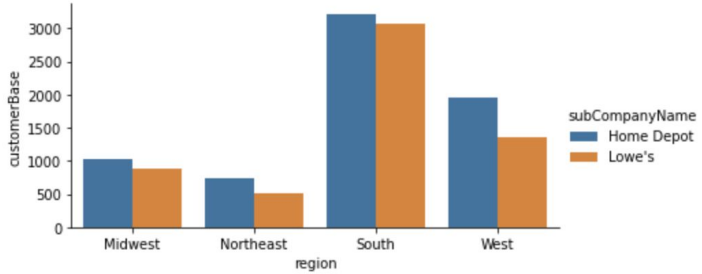
Home Depot



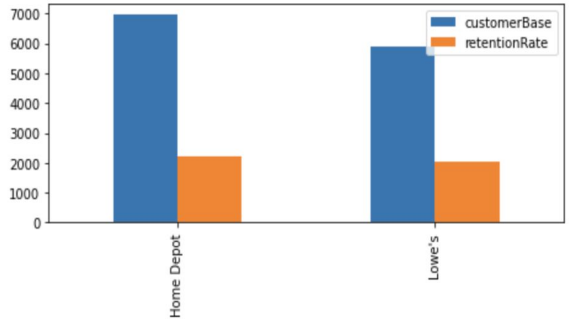
Lowe’s



Popular Vendor

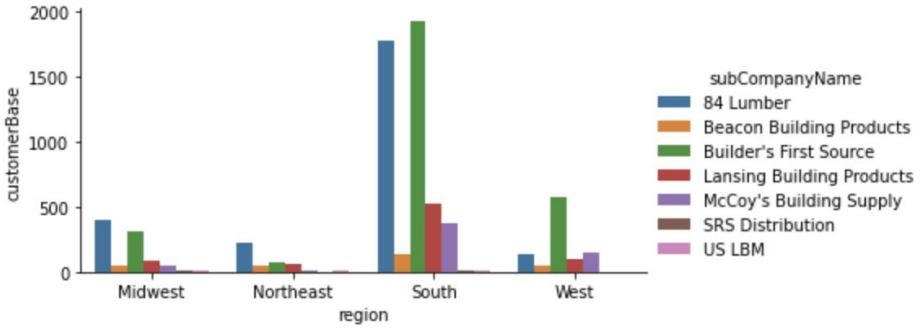


Retention Rate

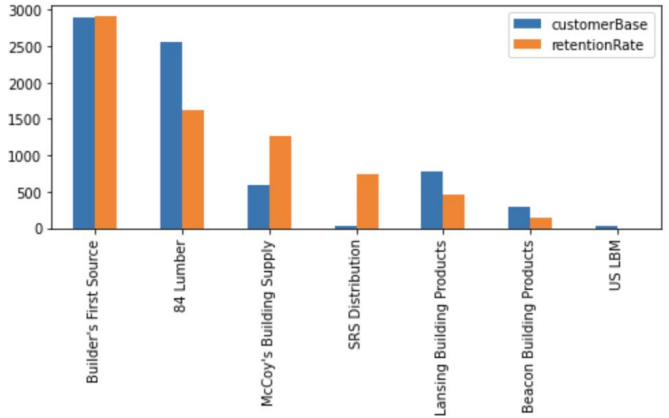


Popular Vendor - Building Material & Lumber

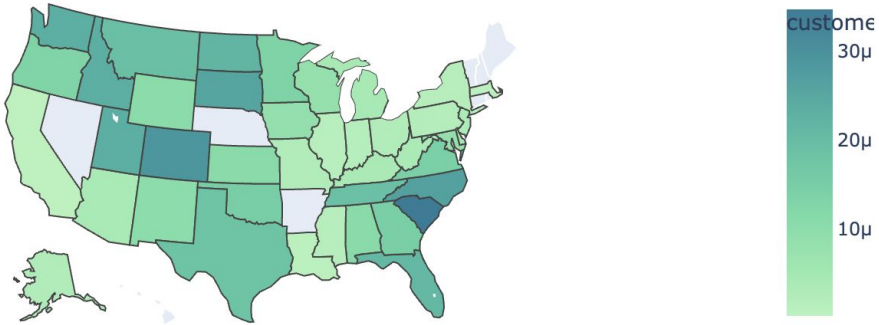
Customer Base



Retention Rate

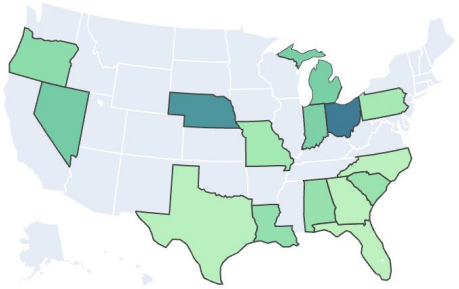


Builder's First Source

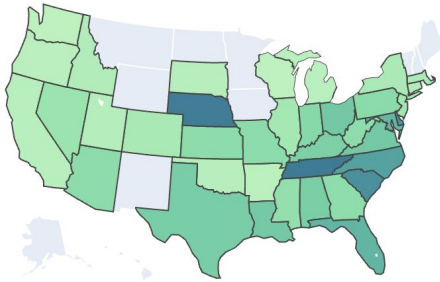


Customer Index - Building Material & Lumber

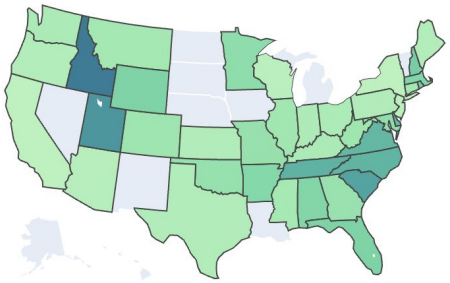
SRS Distribution



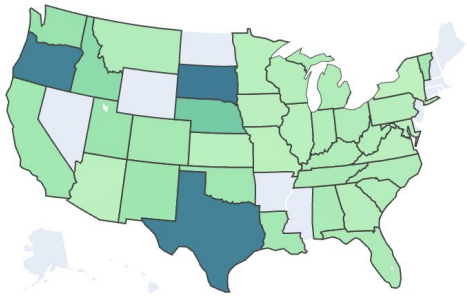
84 Lumber



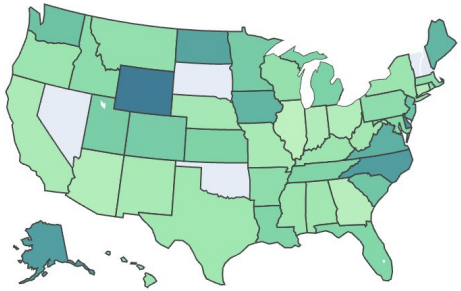
Lansing Building Products



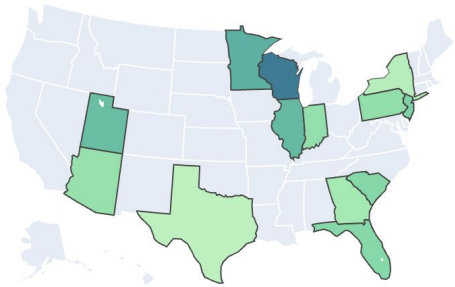
McCoy's Building Supply



Beacon Building Products



US LBM



Data Preprocessing

Predict

	vendor	state	ProjectType	Category	population	month_year	housing_starts	Permits	customerCount
0	84 Lumber	AZ	Remodels	Building Materials and Lumber	7151502.0	2019-01	19.4	2885.0	3
1	84 Lumber	FL	New Homes	Building Materials and Lumber	21538187.0	2019-01	51.4	11299.0	29313
2	84 Lumber	GA	New Homes	Building Materials and Lumber	10711908.0	2019-01	51.4	3876.0	357
3	84 Lumber	MD	New Homes	Building Materials and Lumber	6177224.0	2019-01	51.4	1695.0	200
4	84 Lumber	MD	Remodels	Building Materials and Lumber	6177224.0	2019-01	51.4	1695.0	16
...	...	...	...	...	...	...	...	...	...
20433	Winsupply Inc.	SC	Commercial	Plumbing	5190705.0	2021-12	66.0	3225.0	5
20434	Winsupply Inc.	SC	New Homes	Plumbing	5190705.0	2021-12	66.0	3225.0	1
20435	Winsupply Inc.	TX	Commercial	Plumbing	29527941.0	2021-12	66.0	20315.0	18

Internal Features - Vendor, State, Project type, Category, Month, Year

External Features - Population, Building Permits, Housing Starts

Prediction Model

AutoGluon

- Problem type - Regression
- Train-test split - 0.2
- Train size = (16350, 9)
- Test size = (4088, 9)

Models used by AutoGluon

- LightGBM
- LightGBMXT
- CatBoost
- XGBoost
- NeuralNetMXNet
- LightGBMLarge
- TextNeuralNetwork

Metrics

Train Data

- Root_mean_squared_error : 473.02
- Mean_absolute_error : 178.31
- Explained_variance_score : 0.92
- R2_score : 0.926
- Pearson_correlation : 0.96
- Mean_squared_error : 223746.96
- Median_absolute_error : 69.41

Test Data

- Root_mean_squared_error : 633.57
- Mean_absolute_error : 236.05
- Explained_variance_score : 0.83
- R2_score : 0.834
- Pearson_correlation : 0.92
- Mean_squared_error : 401414.51
- Median_absolute_error : 86.17

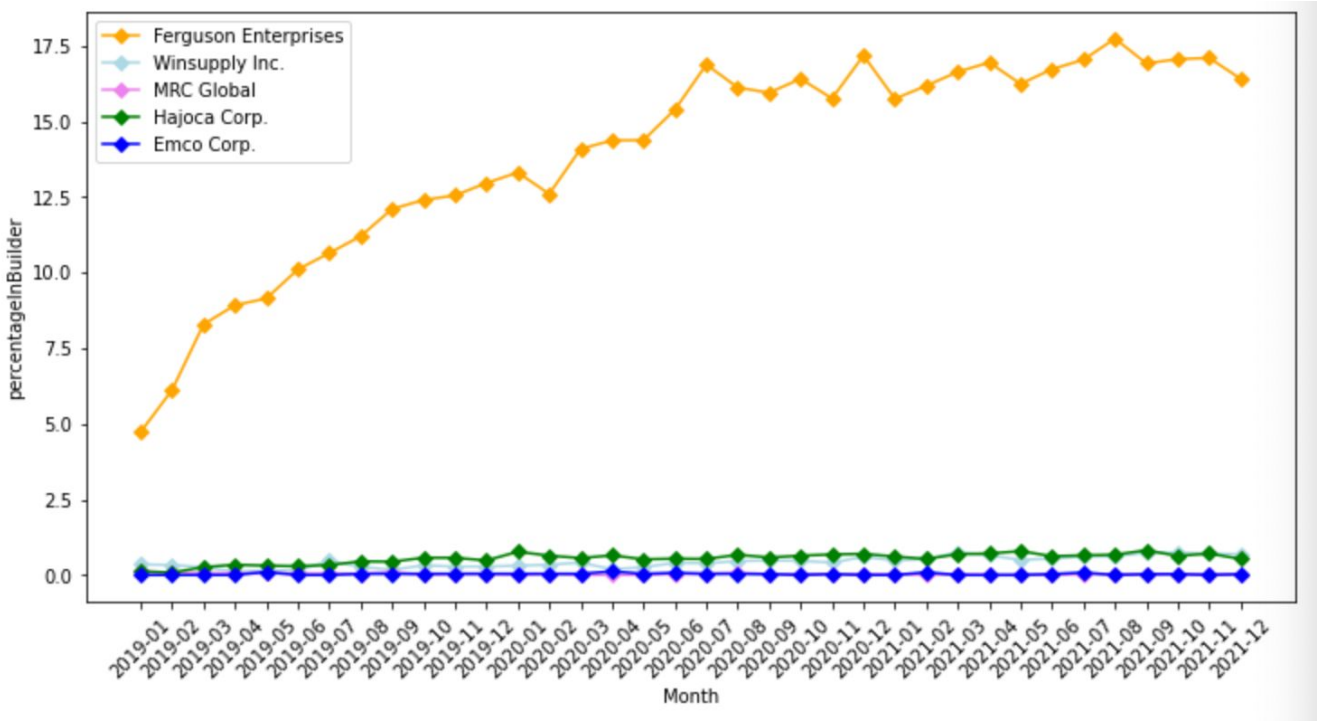
Results

- In the AutoGluon model, we used different evaluation metrics, train-test split approaches, and input datasets to train the model.
- The result shows that the prediction is pretty accurate for a set of inputs while it's not the case with all the inputs.
- With the current datasets the prediction is not stable enough to be implemented in the business.
- But it is possible that this model works very well for a certain customer count cap.

Time Series Analysis

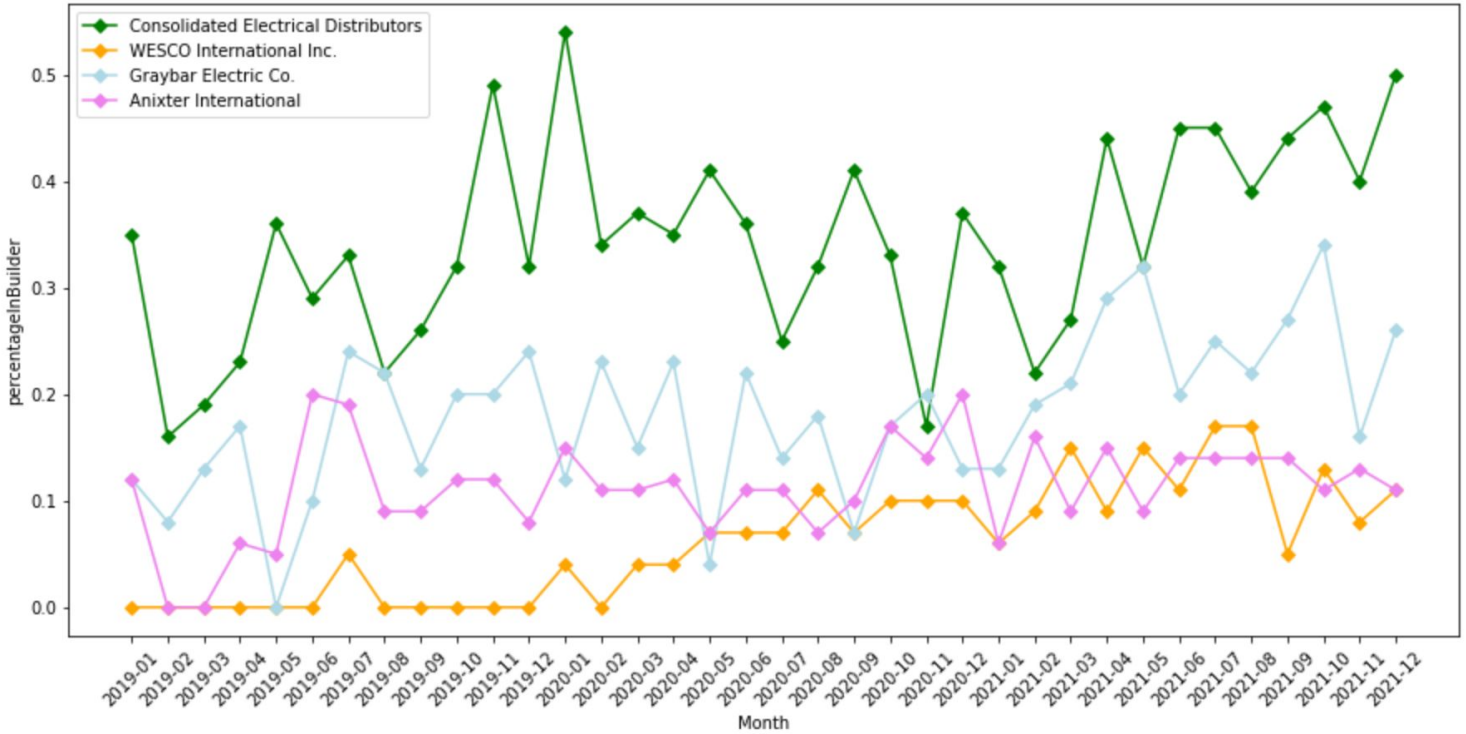
- Identified total customer base of each vendor statewise. Also identified the changes in customer base over time.
    - **Customer base change = number of unique builders who choose a vendor in a month / number of unique builders in the month**

Customer Base Comparison in the Same Category



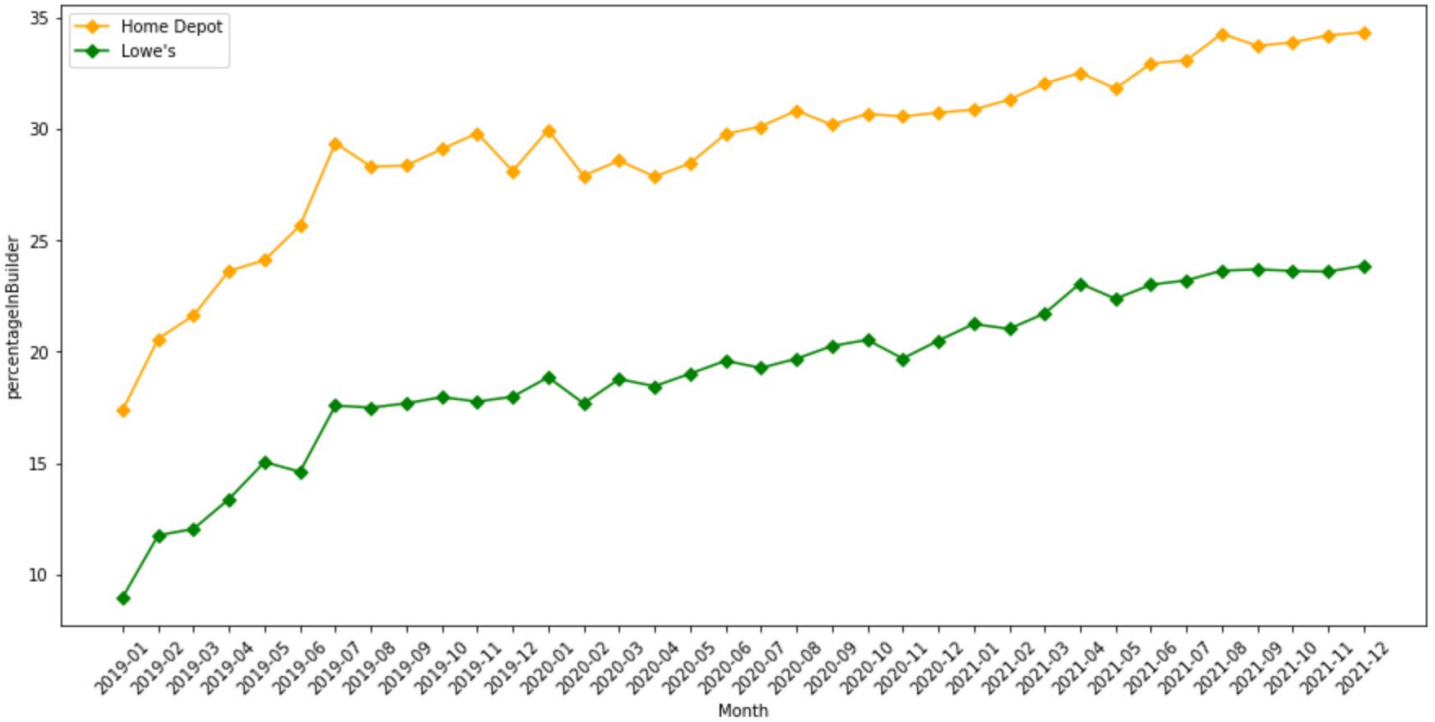
Plumbing

Customer Base Comparison in the Same Category



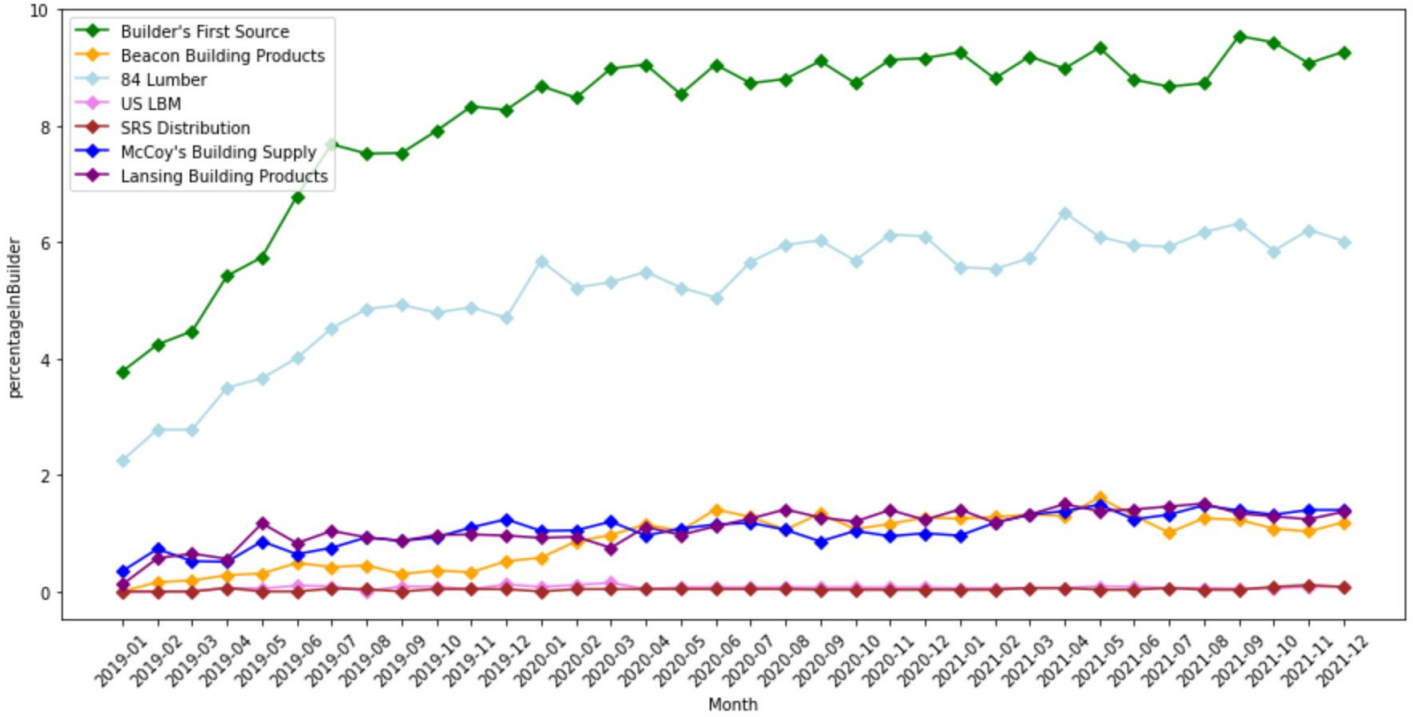
Electrical

Customer Base Comparison in the Same Category



Home Center

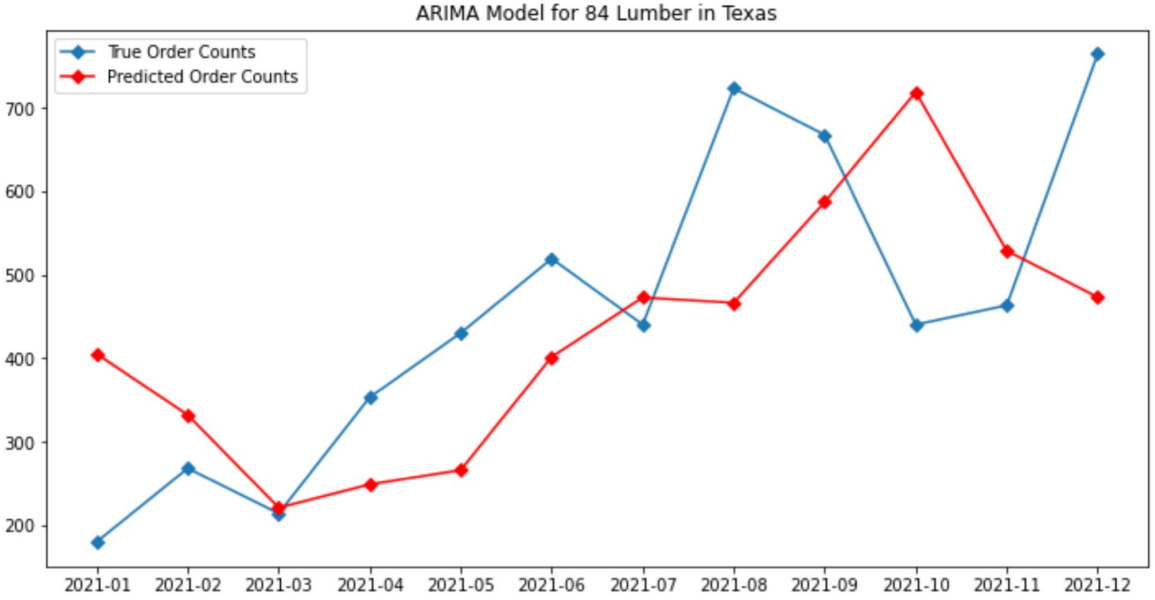
Customer Base Comparison in the Same Category



Building Materials & Lumber

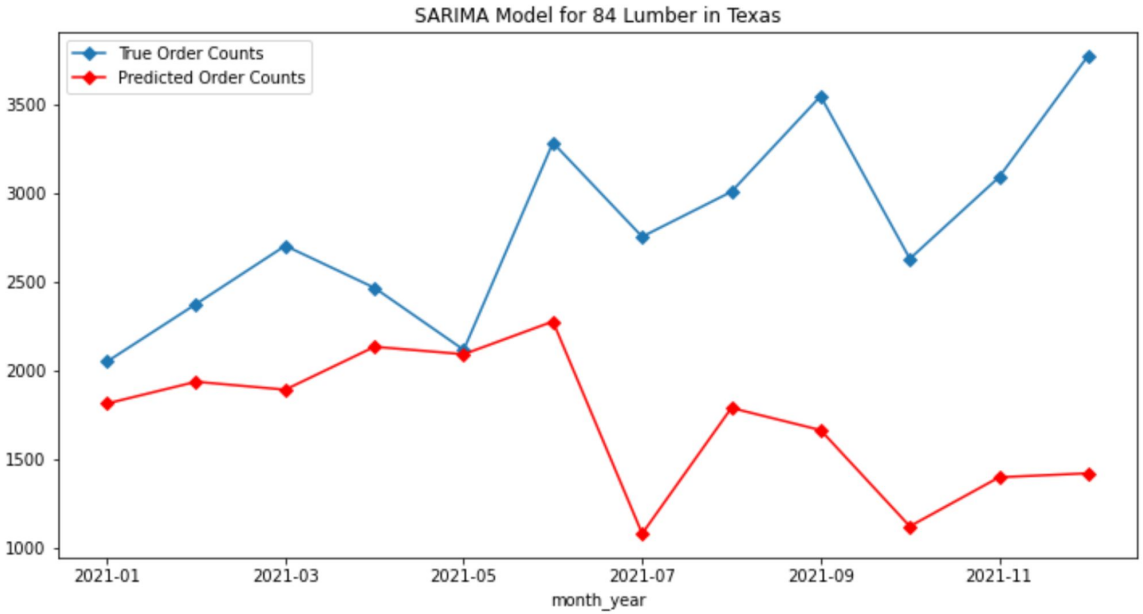
Time Series Forecasting - Order Counts - ARIMA

Since different vendors in different states have very different customer base, we should build model for each vendor in each state



84 Lumber in Texas:
Use data from 2019-01 to 2020-12 to predict order counts from 2021-01 to 2021-12
Test RMSE: 169.882
Test R2: 0.127

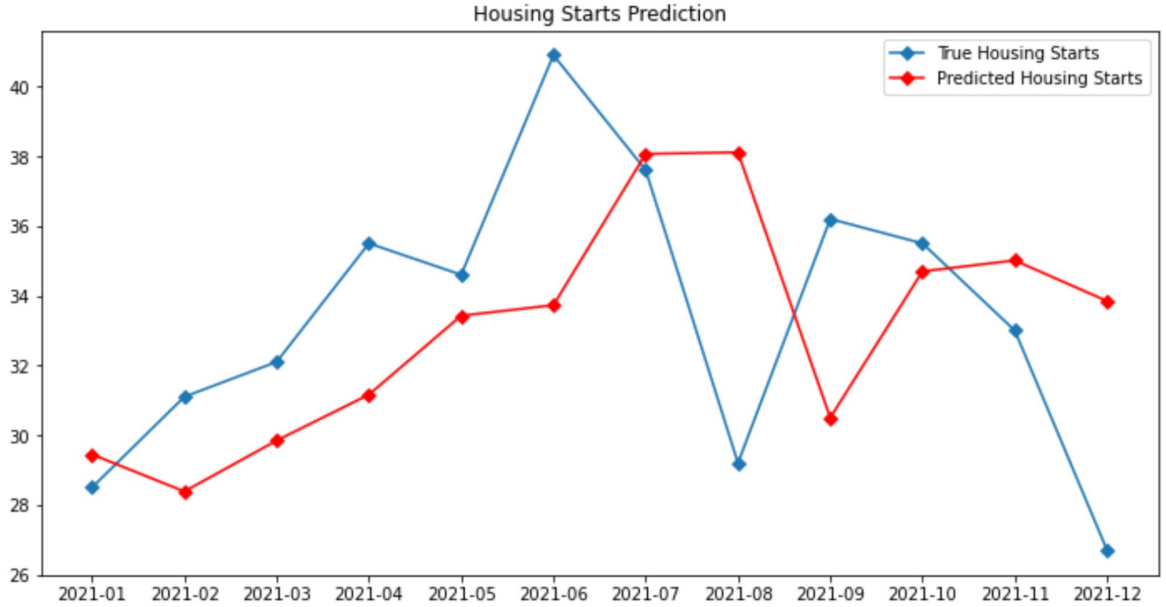
Time Series Forecasting - Order Counts - SARIMA



84 Lumber in Texas:
Use data from 2019-01 to 2020-12 to predict order counts from 2021-01 to 2021-12
Test RMSE: 1306.86
Test R2 score: -5.36

Time Series Forecasting - ARIMA

Build model for each region



Housing Starts in West Region:
Use data from 2019-01 to 2020-12 to predict Housing Starts from 2021-01 to 2021-12
Test RMSE: 4.591
Test R2 score: -0.362

Limitations & Next Steps

Limitations:

- The housing starts and permits datasets from Census Bureau only include new construction, and Census Bureau stopped collecting the equivalent datasets for old buildings, but the builder jobs include both kinds of constructions. This may also influence the prediction.
- The Housing Starts and Permits datasets are lagging. We can only get the data for the current month one month later. So using the monthly Housing Starts or Permits data to predict the customer counts of the same month is not quite reasonable. In the future we plan to use the monthly datasets to predict the customer counts two month later.
- The datasets we already have are not enough to build a reliable model to predict the customer counts.

Limitations & Next Steps

Next Steps:

- Include more external datasets.
- Split the datasets into multiple pieces based on different rules, such as States, regions, vendors to find if the model is reliable for a certain customer count cap.
- Collect more observations to see if time series forecasting works.
- Focus on each region and evaluate which states the company has scope to grow in.
- Consider looking at historical external data to see how the housing market has been affected during times of turmoil.