Buildertrend Final Summary Report

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

About Buildertrend

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.

Project Scope

- The project mainly focuses 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

Work Plan and Timeline

Phase 1: End of March 2022

- Explore the 10 datasets that Buildertrend had provided
- Based on exploration and research into the industry, formulate potential research options
- Draft proposal with scope and objectives

Phase 2: Mid April 2022

- Developed 2 tasks to be accomplished throughout the project
- Cleaning all internal data to make it ready for machine learning models

Phase 3: Mid May 2022

- Exploratory data analysis, data cleaning and wrangling
- Exploring external datasets from the US Census Bureau to be used in conjunction with the internal datasets provided

Phase 4: End of May 2022

- Implementation of ML models on both tasks
- Insights and conclusion

Benefits to Buildertrend:

- Forecasting the number of jobs that the company has 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.

Data Cleaning:

1. **Internal datasets:**
    - Samplejobs - Data entered by builders into the Buildertrend platform that gave information on each of the jobs that they had, including the location, start and end time, the project type, and approximate price.
    - Schedules - More in-depth information on each of the jobs that builders had to show the timeline of each job
    - Subs - Details about Buildertrend's customers' vendors for their orders.
    - purchaseOrderItems - Data entered by builders that briefly records the order they purchase during the construction.
    - purchaseOrderLineItems - the detailed items information for the purchaseOrderItems.

2. **External datasets, all by the US Census Bureau:**
    - Population - The population in the US, divided into 4 different regions, from 2019-2021
    - Median Income - The mean income of different household types within each region from 2019-2021
    - Housing Starts - Census-provided data on the start time of each construction project across the country, itemized by region and time.
    - Building Permits - Census-provided data on the total building permits across the country, itemized by time and series.

Task - 1

Exploratory Data Analysis:

In our initial EDA, we found that while Buildertrend has jobs across all types of residential and commercial construction, by far their highest project type was in ‘new home construction.’ We decided to use this information and focus our analysis on new home construction moving forward, as it’s the one that the company was most interested in, and with the most scope of taking it forward. Additionally, we found that Buildertrend’s jobs were overwhelmingly concentrated in 3 states: Texas, California, and Florida. After discussing these initial insights with the company, we realized that one of the most important things to find would be the states in which they have the highest opportunities moving forward - states that they had not tapped as much. However, we could not pursue this route much further, since the external data that we had was region-wise, and not granular enough to the state level.

These insights gave us a better understanding of the ML model that we had to formulate. After these insights, we decided to look at certain data that the US Census Bureau provides, to understand the demographic that Buildertrend needs to target next. This included US housing starts, population growth, median household income, and household types.

Prediction:

Our primary goal with machine learning was to forecast future jobs that Buildertrend will get, using past Buildertrend data and Census data. To do so, we wanted to use a regression model, and so we decided to use an Autogluon model on the data. This had a few benefits:

- Reads in both text and integer data types
- Runs 7-8 models simultaneously without having to run extra lines of code.

Our final model showed that there is a strong relationship between all the features and our output - the company can use the external data provided by the US Census Bureau to forecast the jobs they will be getting in the future, region-wise.

However, one main limitation of the model was that because the data was only over 3 years, we cannot use it to accurately predict any major volatility in the markets or in the construction industry. So while the company can use the model to predict future jobs, if there are any major changes (e.g. inflation continues to rise, pandemic, market crashes), then the model will not work.

Future Work:

There are a few next steps in this project:

- Conduct time series analysis on the data to see if over time the trends have changed and to see if seasonality is a big determinant of number of jobs
- 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, and try to forecast Buildertrend's business during future difficult periods as well.

Task - 2

Exploratory Data Analysis:

In order to predict the customer base that the partnered vendors would have in the future, we need to identify the features first. The following are the exploratory analysis made on the subs, purchaseOrders and purchaseOrderLineItems datasets to get interesting insights.

- 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
- Calculated a retention rate metric for each vendor based on total number of orders and customer base. A retention rate could simply give the ratio of regular customers of vendors.
  - Retention Rate = Total number of orders for a vendor / customer base of vendor
- Calculated 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
- Implemented geo plots to compare and visualize customer base and customer index for each vendor statewise.
- Found the most popular vendor region wise for each category. The different categories include plumbing, electrical, electrical & plumbing, home center, building materials & lumber. The regions include northeast, west, south, midwest.

Prediction:

The main goal of this task is to predict the customer base that the partnered vendors would have in the future. We used different time series forecasting models and regression models. We used ARIMA and SARIMA models for time series forecasting. The former forecasts the customer base of each vendor whereas the latter forecasts the housing starts. We tried ML models like the Random Forest and Decision Tree regressors for the customer base prediction. Since we didn't get good results, we trained our data on a multimodel ML model. We used AutoGluon because this model would run 7-8 regression models for our problem before predicting a result. So the result could be more reliable. We gave the following features to our model.

- Buildertrend features - project type, category, month, year
- External features - population, building starts, building permits

In the AutoGluon model, we used different evaluation metrics, train-test split approaches, and input datasets to train the model. The result shows that with the current datasets the prediction is not stable enough to be implemented in the Buildertrend business. The prediction is pretty accurate for a set of rows whereas the prediction is not acceptable for the other part. It is possible that this model works very well for a certain customer count cap.

Future Work:

- The datasets we already have are not enough to build a reliable model to predict the customer counts, in the future we will include more external datasets.
- 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.
- We will 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.