A Novel Application of the Bootstrap for New Store Sales Forecasting Containing Sparse Demand Products

Snehalatha Doddigarla, Srishti Bisen, Sukhsagar Jaiswal, Vikas Das, Matthew A. Lanham
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
sdoddiga@purdue.edu; sbisen@purdue.edu; jaiswal5@purdue.edu; das104@purdue.edu;
lanhamm@purdue.edu

ABSTRACT

We provide a bootstrapped-inspired design to both cluster and predict the performance of newly opened stores for a national automotive aftermarket retailer carrying sparse demand products. Our solution could be successfully implemented by other retailers that are not as forecast-accuracy-constrained as we were in our problem. Our approach to this problem was to first capture the variability of any store's forecast versus their actual sales. We assume the retailer will only generate an assortment they believe to be optimal. Next, we identify similar stores based on the new store's demographic and store characteristic profiles using multiple clustering techniques and store these sets. Using these different store sets, we develop store forecasts for each store using several different prediction approaches. These models showed varying statistical performance and varying store sales forecasts. For each cluster-group/predictive-model-group, we average the prediction and run this several times and store the results for each store. Summarizing the bootstrapped outcomes provided the merchant team an expected sales forecast distribution, which helped to not only to give a forecast point estimate but also a confidence interval for better planning and risk assessment for the new store. This design allowed the retailer to 1) estimate performance (point estimate) and risk (forecast distribution), and 2) provide a means to estimate when the new store would likely break-even from their initial investment.

Keywords: Bootstrap, Sales Forecast, Predictive Analytics, PAM, Jaccard's coefficient, MAPE, Cluster validation

INTRODUCTION

Strategic planning based on dependable sales forecasts is an essential measure for successful business management, especially for a market-oriented industry like an automotive aftermarket. Forecasting goes beyond the financial analysis, and it assists in setting the goals for the future. It helps in answering questions about footfalls, operations, supply, and customer spending for a given period. It also serves as a guide for internal expenses to maintain functioning, manufacturing, marketing, and administration.

Return on investment is the most crucial factor while choosing a new site, and sales forecasting for a potential new site becomes essential for a new location planning decision. How precise can sales be predicted?

Sales forecasts cannot only be based on intuitive guesses. Mathematical models, along with powerful computer resources, are crucial for obtaining accurate predictions. Methods based on statistical learning theory are potent instruments to get insight into internal relationships within massive empirical datasets, producing reliable and even highly accurate forecasts. There have been continuous efforts to develop and improve forecasting models.

In this paper, we review the research as it applies to retail forecasting, drawing boundaries around the field to focus on the automotive aftermarket industry. We also discuss various improved sales forecast methodologies and models for the automobile market that can deliver accurate predictions and, at the same time, maintain the ability to explain the underlying model.

The first part of our research describes the methods used to form clusters of similar stores for the various regions. We establish clusters /groups of items based on store demographics, vehicles in operation near the stores, and other similarity measures and then compute a standard forecast for each cluster of objects. We use different mechanisms for clustering like K-means, PAM, Clara, Hierarchical, Spectral clustering. The primary methods used to find the optimal number of clusters is based on cluster validation and bootstrap evaluation of cluster. These methods allow us to find the optimal number of clusters based on metrics such as Dunn, Connectivity, AD, and Jaccard Coefficient.

The second part of our research entails using bootstrapped Time-series methods to gain the best possible forecasts. Time-series forecasting is one of the most widely applied techniques for accurate predictions in business, used extensively in finance, supply chain management, and in production and inventory planning. We employ time series analysis along with traditional data mining procedures, like regression, decomposition, and other adaptive methods to capture patterns in data over time and project the founded patterns in prospect, with the assumption that the trends will repeat themselves in the future. Seasonality is a significant concern when it comes to model time series forecasting for automotive industries as they have some definite seasonal trends with peak demands occurring in spring and fall usually.

We employ classic time series models like ETS (Error Trend Seasonality), Holt-Winters Exponential Smoothing, and ARIMA (Autoregressive Integrated models of Moving Averages) to forecast sales of clusters, bootstrap the forecasts for n number of times and calculate the average and then take their mean as the final forecasted sales for higher accuracy. Further, we check the accuracy of our time series models with measures like MAPE (Mean Absolute Percentage Error), MASE (Mean Absolute Scaled Error), MAE (Mean Absolute Error) and RMSE (Root Mean Square Error). Based on the sales forecasts projected by our model, we then make recommendations to our client to optimize their supply chain operations and efficiently plan the operations of the new store.

The remaining paper is organized in the following manner: a literature review of the existing related articles and papers by reputable sources, the next section introduces data that we have used to build our analysis and model. The third section provides insight into our analytical approach and solution that will assist the client in selecting the location to get a good return on investment. This section also describes the forecasting model and statistical error measures. Section four discusses the findings and results, and the next section focuses on limitations, further research avenues, and a conclusion.

LITERATURE REVIEW

The research on sales forecasts for a new store based on the past sales history of similar stores has not been done extensively. It needs to be explored in many ways. Demand and sales forecast for exiting products and stores has been a norm in research, but research for the same for a new store is limited.

Retailers face forecasting problems in a wide array of areas from strategic to operational, including forecasting for competing channels, for various products of a brand, aggregated at multiple levels. Fildes, Roberts, Shaohui, and Stephan (2019) demonstrated this in their study which is mainly focused to understand the sales forecast at the store level. Store sales forecasts can be of two types: forecasts for existing stores and new stores. Sometimes, experience and intuition are leveraged to gauge the sales using a checklist, as is the case of a judgment or econometric models. Analog regression models generate turnover forecasts for a new store, considering the store characteristics, data on competition, store accessibility, and the local market. There has been evidence that such models have seen success in considering 80% of store turnover. Regression models are highly explainable, but the use of the regression equation is generally insufficient to predict the potential performance of the new store.

One of the earlier approaches for this was the Analog approach for new sales forecast has been discussed by Rogers and Greens (1979). An analog engine, commonly referred to as a 'peer group comparison,' forecasts sales of new stores by comparing the new store in question to similar existing stores and evaluate the new store's performance analogous to most similar stores. Analog engines are not predictive in nature but play a crucial role in interpreting regression or time series-based model for sales forecast. One thing to keep in mind is that the forecasting models are aggressive because they are forecasting sales for a new store in a new market. In contrast, the analog engines are conservative and reflect the analogous stores' state in the current market.

The paper which closely relates to our current study is Li Na, Cai Ying, and Bi De's article (2019), which explores the empirical analysis for a store sales forecast on multidimensional data. ARIMA and ARMA model was extensively used by the authors to predict the sales by including the input factors such as time dimension, geographical dimensions, and promotional methods. The research assisted us in formulating an idea about adding the demographics into the model to forecast the sales.

K-mean clustering is one of the most popular clustering techniques and the most widely used partitioning clustering methods. There are certain drawbacks to using this method is that it suffers from sensitivity to outliers, which deforms the distribution of data due to significant values in the outliers. The paper by Faisal Abid (2014) discusses in detail the advantages and drawbacks of the K-means and K-medoids method. K-medoid has various evolved algorithms that are addressed in this paper, such as PAM (Partitioning around Medoids) and CLARA (Clustering large applications based upon Randomized search).

Another article by Christian Henning (2007) provides more insight into the assessment of cluster stability. It is to be noted that stability is not the only measure to validate a cluster as a stable cluster could be meaningless as well. Some alternative methods of cluster validation are homogeneity or separation-based validation indexes, comparison of different clustering methods on the same data, visual cluster validation, tests of correlation of the data set against a clustering alternative and use of external information, see Gordon (1999), Haldiki (2002), Hennig (2005), and Milligan and Cooper (1985).

Jaccard coefficient is calculated for clustering methods and different bootstrapping methods. Jaccard coefficient measures the similarity between two subsets of a set based on set membership. Formally the Jaccard coefficient is defined as an intersection over the union of sets. It can be stated as:

$\gamma(C, D) = |C \cap D| |C \cup D| , C, D \subseteq xn$.

Median, a trimmed mean or the number of dissolutions or Jaccard coefficients greater than 0.75, indicates that the cluster is stable. The bootstrapping methods explored here are jitter, sub-setting, replacing points by noise, and jitter-boot, which were used in our design to check the stability of a valid cluster.

There have been other methods which include Hybrid ARIMA with other techniques such as Neural Network Model. ARIMA (Autoregressive Integrated Moving Average) and ANN (Artificial Neural Network) have been two of the most popular linear and nonlinear models, respectively, for time series forecasting. The paper discusses a hybrid methodology combining both these models in a way that improves the forecasting accuracy achieved by the models when used separately. The motivation of a hybrid model comes from the fact that real-world time series are seldom pure linear or nonlinear in nature, in which case pure ARIMA or ANN models fail to capture the pattern. Statistical, as well as empirical pieces of evidence in the paper, suggest that using a hybrid model, comprising of models that intensely disagree with each other, has lower generalization error as well as lower model uncertainty.

Hulsmann, Reith, and Fredrich's (2012) approach for building a general forecast model relates to our research to a certain degree. This paper expands on previous contributions to the field of vehicle registration forecasting. Previous time series models with a trend estimation by Multivariate Linear Regression (MLR) and Support Vector Machine (SVM) have produced reliable forecasts. In this paper, high importance was given to building and testing models with good accuracy and explicability for two countries at a yearly, quarterly, and monthly level. Data Mining methods that were tested include Decision Trees, K-nearest-Neighbors, Random Forest, Ordinary Least Square, Quantile Regression were used. The use of unlimited exogenous data, instead of a mixture of absolute and relative data used with Decision Trees, delivered meaningful results with high accuracy.

Customer convenience is of utmost importance in the retail industry. Factors like store location can have a significant impact on business and necessarily make or break your business. Over the years, hypermarket retail store activity has significantly slowed down with small or medium markets with more strong dynamism. Spending on consumer proximity by investing in stores of more modest dimensions is strategically a long-term better decision than spending on the quality of goods and services. Retailers have always known that location is important; however, understanding all aspects of store performance, site potential, and consumer behavior demands more information like the demographics, socio-economic, and competitor data. In this context, the analysis and research are given by Steve Wood and Andrew Tusjer (2007), shows that the demand forecasting becomes paramount in enabling the business to make viable decisions regarding store location. For example, Tesco, which is one of the largest retailers in the United Kingdom, invested in forecasting projects to improve the store turnover forecast with only a 10% mean error starting from simple regression models to gravitational models resulting in substantial commercial success to them. Similarly, they performed store location analysis to gain competitive advantage using analog based procedures, gravitational models, analog-based regression models, and discriminant analysis.

One of the disadvantages of using a simple time series forecasting is that even though it considers demographic measures like population into account, however, it loses on other age-dependent factors like infant mortality rates, warranty expiration, wear-out, vehicle attrition and retirement, periodic maintenance and seasonal mileage. Laurence Stefan's (2004) research in actuarial forecast discusses this. An actuarial forecast breaks a simple forecast into its actuarial demand rate component and count of vehicles in different forecast intervals, which is split across by age. Actuarial demand rates are probabilities of failure in the various age intervals given there is no failure before any age intervals. One of the key elements for developing an actuarial forecast is the actuarial demand rate, which now is not readily available. Hence, the expansion of this methodology is constrained.

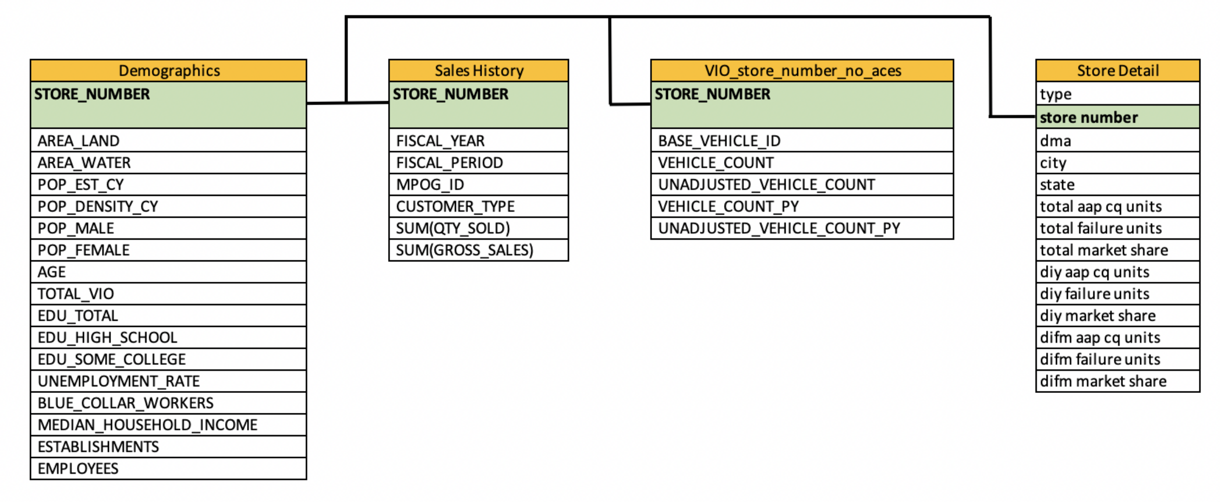
Put the problem is how can we make actuarial forecasts if you don't know next year's actuarial rate. If we have complete data, simple linear regression can provide adequate extrapolation, albeit with modeling uncertainty. There are two key advantages to actuarial forecast than simple forecast is that it considers the entire population data, which reduces the risk associated when we use sampling. The second key advantage is that missing data does not limit it. The estimation-maximization method, which is used to estimate the actuarial demand rate works even when the data is missing.

Although various research has been done for similar business problems, further research will help a business plan better to achieve new growth in sales and expansion. Our study provides a bootstrapped forecast model that assumes that stores in a cluster behave similarly in terms of sales. We forecast the sales for a cluster that uses a bootstrapped time series data. Our study, in addition to forecasting sales for a new store based on similar stores' sales history, it also provides a means to estimate break-even point, and predict risk, and this makes our study novel.

DATA

Internal Data:

There were 12 tables made available to us by the client on MariaDB database out of which we have used the following tables for our model building and analysis – Demographics, Sales_History, VIO_store_number_no_aces and Store_Detail. We merged these tables on STORE_NUMBER as the primary key. The following schema demonstrates the tables and their relationships visually:



Demographics
STORE_NUMBER
AREA_LAND
AREA_WATER
POP_EST_CY
POP_DENSITY_CY
POP_MALE
POP_FEMALE
AGE
TOTAL_VIO
EDU_TOTAL
EDU_HIGH_SCHOOL
EDU_SOME_COLLEGE
UNEMPLOYMENT_RATE
BLUE_COLLAR_WORKERS
MEDIAN_HOUSEHOLD_INCOME
ESTABLISHMENTS
EMPLOYEES

Sales History
STORE_NUMBER
FISCAL_YEAR
FISCAL_PERIOD
MPOG_ID
CUSTOMER_TYPE
SUM(QTY_SOLD)
SUM(GROSS_SALES)

VIO_store_number_no_aces
STORE_NUMBER
BASE_VEHICLE_ID
VEHICLE_COUNT
UNADJUSTED_VEHICLE_COUNT
VEHICLE_COUNT_PY
UNADJUSTED_VEHICLE_COUNT_PY

Store Detail
type
store number
dma
city
state
total aap cq units
total failure units
total market share
diy aap cq units
diy failure units
diy market share
difm aap cq units
difm failure units
difm market share

Fig 1. Schema of tables

To predict the sales of new stores, we will cluster similar stores together and predict the mean sales of each cluster. In order to cluster similar stores together, we have taken parameters like demographics (area of land near the store, area of water, population density, population of males, total education of population, unemployment rate, number of employees etc.), vehicle count information and store details (designated market area, city, state, type of store, failure units etc.) in our model.

Based on the clustering results, we train an k-NN model to map test data stores to the respective stores. We train the model on all the variables used in the clustering model along with the clusters the train data set stores are assigned to. We then forecast the sales of these test data stores using Time Series by taking variables fiscal_year, fiscal_period and sum(gross_sales).

The data dictionary below describes each variable used by in our models:

Variable name	**Variable type**	**Description**
Table: Demographics		
STORE_NUMBER	Numeric	AAP Store Number
AREA_LAND	Numeric	Area of census tract that is land
AREA_WATER	Numeric	Area of census tract that is water
POP_EST_CY	Numeric	Estimated population current year
POP_DENSITY_CY	Numeric	Population density current year
POP_MALE	Numeric	Male Population
POP_FEMALE	Numeric	Female Population
AGE	Numeric	Average age
TOTAL_VIO	Numeric	Total number of vehicles in operation
EDU_TOTAL	Numeric	Population with at least some education
EDU_HIGH_SCHOOL	Numeric	Number of people with high school education as maximum education
EDU_SOME_COLLEGE	Numeric	Number of people with some college education as maximum education
UNEMPLOYMENT_RATE	Numeric	Local average unemployment rate
BLUE_COLLAR_WORKERS	Numeric	Number of blue-collar workers
MEDIAN_HOUSEHOLD_INCOME	Numeric	Median household income ($)
ESTABLISHMENTS	Numeric	Number of locations where business is conducted
EMPLOYEES	Numeric	Number of paid employees working in the area
Table: Sales_History		
STORE_NUMBER	Numeric	AAP Store Number
FISCAL_YEAR	Numeric	Year
FISCAL_PERIOD	Numeric	Period (13 per year with 4 weeks per period)
MPOG_ID	Numeric	Product category for part types
CUSTOMER_TYPE	Categorical	Type of sale (DIY, DIFM, Other)
SUM(QTY_SOLD)	Numeric	Sum of quantity sold in that period
SUM(GROSS_SALES)	Numeric	Sum of gross sales in that period
Table: VIO_store_number_no_aces		
Variable name	Variable type	Description
STORE_NUMBER	Numeric	Store number
BASE_VEHICLE_ID	Numeric	Primary Key
VEHICLE_COUNT	Numeric	Number of registered cars attributed to store by weighting surrounding census tracts by store market share
UNADJUSTED_VEHICLE_COUNT	Numeric	Number of registered cars in the census tract without any weights or adjustments

VEHICLE_COUNT_PY	Numeric	Past year number of registered cars attributed to store by weighting surrounding census tracts by store market share
UNADJUSTED_VEHICLE_COUNT_PY	Numeric	Past Year number of registered cars in the census tract without any weights or adjustments
Table: Store_Details		
Type	Categorical	AAP or CQ store
store number	Numeric	Store number
Dma	Categorical	Designated Market Area - area that receives the same television and radio station ads
City	Categorical	Store location
State	Categorical	Store location
total aap cq units	Numeric	Number of units the store has sold annually
total failure units	Numeric	Total possible sku sales in surrounding area based on weighting of nearby census tracts
total market share	Numeric	Percent of market served by AAP sales calculated from total units and total failure units
diy aap cq units	Numeric	Number of units classified as DIY as a subset of total aap cq units. DIY = do it yourself
diy failure units	Numeric	Total number of DIY units sold in the census tract not limited to sales from AAP/CQ stores
diy market share	Numeric	Percent of DIY market served by AAP sales calculated (diy aap cq units / diy failure units)
difm aap cq units	Numeric	Number of units classified as DIFM as a subset of total aap cq units. DIFM (do it for me) iincludes sales to businesses or professional mechanics
difm failure units	Numeric	Total number of DIFM units sold in the area not limited to sales from AAP/CQ stores
difm market share	Numeric	Percent of DIFM market served by AAP sales calculated (difm aap cq units / difm failure units)

Table 1. Data Dictionary of tables

External data:

We geo encoded the city and state details into latitudes and longitudes in order to reduce the dimensionality of our clustering model. The table 'uscities' contains the same.

The data dictionary below describes each variable used by in our models:

Variable name	Variable type	Description
Table: uscities		
STORE_NUMBER	Numeric	Store number
City	Categorical	City in which store is located
State	Categorical	State in which store is located
Latitude	Numeric	Latitude of the store location
Longitude	Numeric	Longitude of the store location

Table 2. Data used in models

METHODOLOGY

The primary problem focus in our study is to forecast sales for a new store based on similar stores' sales history. In order to align the business problem of estimating the performance and risk into an analytical problem, we broke the problem down into two sub problems.

1) Clustering of Similar Stores

2) Time Series Forecasting

Our study consists of the following methods and models:

- **Data Preparation**: Merged tables to be used in our analyses on STORE_NUMBER, geo-encoded the city and state details into latitudes and longitudes to reduce the dimensionality of our models, removed highly correlated variables from our models to avoid multicollinearity. We performed correlation and removed the variables with correlation greater than 0.9. The correlation plots are show below:

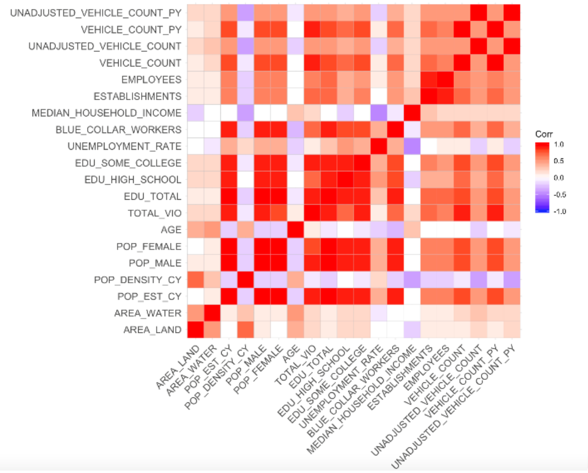


Fig 2. Correlations of all variables

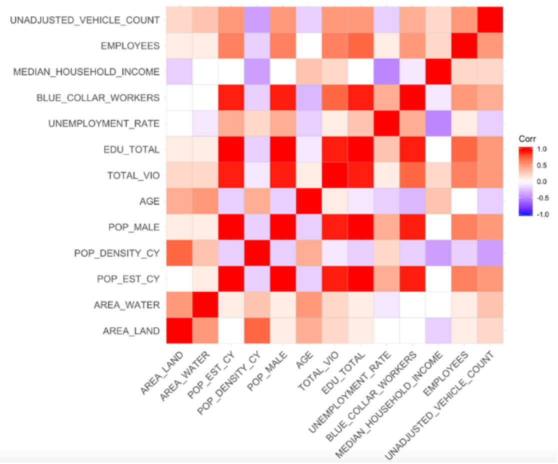


Fig 3. Correlations after variable selection

- **Bootstrapped Clustering**: Bootstrapped results from clustering models like k-means, Clara, PAM, Spectral clustering and Hierarchical clustering to determine the optimal number of clusters for our stores to map to. We took store demographics, details and market share data for clustering. We then checked the efficacy of our bootstrapped clustering result from statistical measures like Jaccard coefficient and Dissolution rate.
- **Cluster Classification with k-NN**: In order to map new stores to clusters, we used k-NN (K-nearest neighbors) for classification. We divided our dataset into train and test set, with test set consisting of details of stores opened in 2017 and train set containing details of all other stores. We trained the k-NN model with all variables used in the clustering methodology along with the cluster assigned to stores as the target of our model. We then used the test set to map the stores to clusters.

- **Sales Data Aggregation**: To predict the sales of each cluster, we began with grouping the sum of sales of each store based on the fiscal period and fiscal year. We then used this aggregated sales data to group the mean of sum of sales of each store within a cluster on fiscal period and fiscal year.
- **Time Series Analysis**: We begin with transforming the sales history data into time series data beginning with the 1st period of 2014 and ending at the 13th period of 2019. We then split our time series data into train and test data as above, with the test data containing the sales history of all stores opened in 2017 for year 2017-2018. We forecast the sales of these stores for year 2018-2019 and check the accuracy of our forecasts with the statistical measure MAPE.

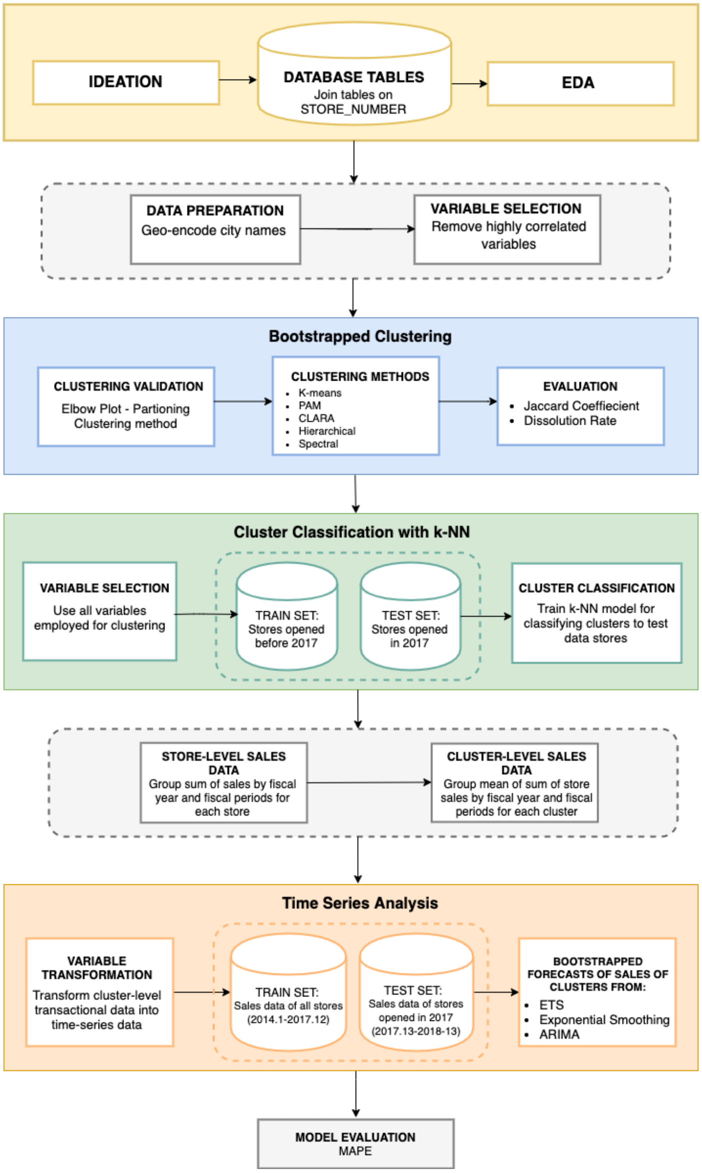

IDEATION
DATABASE TABLES
Join tables on STORE_NUMBER
EDA
DATA PREPARATION
Geo-encode city names
VARIABLE SELECTION
Remove highly correlated variables
Bootstrapped Clustering
CLUSTERING VALIDATION
Elbow Plot - Partioning Clustering method
CLUSTERING METHODS
K-means
PAM
CLARA
Hierarchical
Spectral
EVALUATION
Jaccard Coeffiecient
Dissolution Rate
Cluster Classification with k-NN
VARIABLE SELECTION
Use all variables employed for clustering
TRAIN SET: Stores opened before 2017
TEST SET: Stores opened in 2017
CLUSTER CLASSIFICATION
Train k-NN model for classifying clusters to test data stores
STORE-LEVEL SALES DATA
Group sum of sales by fiscal year and fiscal periods for each store
CLUSTER-LEVEL SALES DATA
Group mean of sum of store sales by fiscal year and fiscal periods for each cluster
Time Series Analysis
VARIABLE TRANSFORMATION
Transform cluster-level transactional data into time-series data
TRAIN SET: Sales data of all stores (2014.1-2017.12)
TEST SET: Sales data of stores opened in 2017 (2017.13-2018-13)
BOOTSTRAPPED FORECASTS OF SALES OF CLUSTERS FROM:
ETS
Exponential Smoothing
ARIMA
MODEL EVALUATION
MAPE

Fig 4. Flow diagram of models

MODELS

Splitting into Test and train datasets:
The validation technique is used for validation. The data was divided into 13 training data records for each store, which were not opened in 2018. The test set for the model is the sales history of all the stores opened in the year 2017. The modeling technique used for predictions is K-means, PAM, CLARA, Hierarchical Clustering, Spectral Clustering, KNN, Exponential Smoothening, Holt-Winters, and ARIMA. Further, each model was fine-tuned to build a more robust model. The following are the details of the models that were experimented with and gave us good results for our problem.

Clustering Methods:
We compared different clustering algorithms, namely K-means, PAM, Clara, Spectral Clustering, and Hierarchical clustering, to get an optimal number of most meaningful, stable clusters.

K-Means:
K-means algorithm is an iterative algorithm that partitions the given data in a predefined number of non-overlapping subgroups, which are called clusters, where each data point belongs to only one subgroup. The algorithm tries to make the data points in a cluster as similar as possible while keeping the clusters as different as possible.
The way the k-means algorithm works is as follows:
1. Specify the number of clusters.
2. Initialize the center called centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep repeating until there is no change to the centroids. i.e., clusters don’t dissolve any further (assignment of data points to clusters isn’t changing).
Pros: It is a simple model, which scales to a large dataset, and generalize to clusters of different shapes and sizes.
Cons: Choosing the number of clusters manually and the outliers have a more significant impact on centroid positioning.

PAM:
PAM stands for “partition around medoids.” The algorithm is intended to find a sequence of objects called medoids that are located centrally in the clusters. Objects that are tentatively defined as medoids are placed into a set S of selected objects. The goal of the algorithm is to minimize the average dissimilarity of objects to their closest next selected object.
The algorithm has two phases:
(i) In the first phase, BUILD, a collection of k objects is selected for an initial set S.
(ii) In the second phase, SWAP, one tries to improve the quality of the clustering by exchanging selected objects with unselected objects.

Pros: It’s intuitive, and more robust to noise and outliers compared to k-means
Cons: It is not suitable for clustering non-spherical and is more time consuming and computer-intensive

CLARA:

CLARA stands for “clustering large application.” The CLARA algorithm is as follows:
1. Create random samples from the original dataset, multiple subsets with a fixed size
2. Compute PAM on each subset and choose the corresponding k medoids and assign each observation of the entire data set to the nearest closest medoid.
3. Calculate the mean (or the sum) of the dissimilarities of the observations to their closest medoid. This is used as a measure of the goodness of the clustering.
4. Retain the sub-dataset for which the mean (or sum) is minimal. Further analysis is carried out on the final partition.

Pros: It is an extension of PAM and is more efficient for massive datasets
Cons: The best medoid may not be selected during initial sampling, and if sampling is biased, we will not get the best clustering

Hierarchical Clustering:

Hierarchical clustering treats each observation as a separate cluster. Then, it executes the following two steps repeatedly:
1. Identify two clusters that are closest together
2. Merge the two most similar clusters.

The steps mentioned above continue until all the clusters are merged.
Hierarchical clustering works by sequentially merging similar clusters, as shown above. This is known as agglomerative hierarchical clustering.
It can also be done by grouping all the observations into one cluster, and then successively splitting these clusters. This is known as divisive hierarchical clustering.
Pros: It is easier to understand and gives dendrogram as an output
Cons: Dendrograms can be easily misinterpreted, and the decisions are arbitrary

Spectral Clustering:

Spectral clustering is a technique with roots in graph theory, where the approach is used to identify nodes in a graph based on the edges connecting them. The method is flexible and allows us to cluster the non-graph data as well.
Spectral clustering uses information from eigenvalues of special matrices built from the graph or the data set. The steps for the algorithm are as follows:
1. Build an adjacency matrix from the graph
2. Create Graph Laplacian (Degree matrix – Adjacency matrix)
3. The eigenvalues of the Laplacian indicate the number of clusters

4. Use Eigenvectors to find the actual cluster labels
Pros: Handles non-convex and arbitrary data with transitive relations
Cons: Computing massive for large datasets for calculating Eigen Vectors and not suitable for noisy datasets

Selection Criteria: Stability, Dunn Index, Connectivity, and AD. PAM gave the best possible combination result of these selection criteria.

KNN:

KNN stands for ‘k-nearest neighbors.’ It is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. We are using classification here to map new stores to existing clusters based on the store parameters.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label.

Pros: It is a simple model with no need to tune parameters and can be used for regression and classification

Cons: The speed of the algorithm is dependent on the number of predictors and the size of data

Timeseries Model:

Holt-Winters:

The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations: one for the level ℓ_t, one for the trend b_t, and one for the seasonal component s_t, with corresponding smoothing parameters α, β^*, and γ.

We are using the additive method and the time series for the same is split into 3 components: value, trend and seasonality. Each of the components are predicted using exponential smoothing with different smoothing coefficients.

Model Formation:

$y_{t+h\mid t} = \ell_t + hb_t + s_{t+h-m(k+1)}$

$\ell_t = \alpha\ (y_t - s_{t-m}) + (1-\alpha)\ (\ell_{t-1} + b_{t-1})$

$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta*)\ b_{t-1}$

$s_t = \gamma\ (y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)\ s_{t-m},$

where k is the integer part of (h−1)/m, which ensures that the estimates of the seasonal indices used for forecasting come from the final year of the sample.

Pros: It is a simple model which can be fit to different data sets and interpretable

Cons: It does not capture Trend and Seasonality which leads to a lower accuracy

ARIMA:

ARIMA models aim to describe the autocorrelations in the data and is based on the idea that the information in the past values of the time series can alone be used to predict the future values.

Model Formation: An autoregressive model of order p can be written as
$y_t = c+\phi_1 y_{t-1}+\phi_2 y_{t-2}+\cdots+\phi_p y_{t-p}+\varepsilon_t$,
where ε_t is white noise. This is like a multiple regression but with lagged values of y_t as predictors. We refer to this as an AR(p) model, an autoregressive model of order p.
Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model.
$y_t=c+\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q}$,
where ε_t is white noise. We refer to this as an MA (q) model, a moving average model of order q. Of course, we do not observe the values of ε_t, so it is not really a regression in the usual sense.
If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. ARIMA is an acronym for AutoRegressive Integrated Moving Average. The full model can be written as $y'_t=c+\phi_1 y'_{t-1}+\cdots+\phi_p y'_{t-p}+\theta_1\varepsilon_{t-1}+\cdots+\theta_q\varepsilon_{t-q}+\varepsilon_t$,
where y'_t is the differenced series.
The “predictors” on the right-hand side include both lagged values of yt and lagged errors. We call this an ARIMA(p,d,q) model,
where, p= order of the autoregressive part, d= degree of first differencing involved, and q= order of the moving average part. The same stationarity and invertibility conditions that are used for autoregressive and moving average models also apply to an ARIMA model.
Pros: It is a simpler and robust univariate model which usually gives more accurate results
Cons: It is sensitive to tuning parameters and less reactive to seasonal components

ETS:
Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. It uses an exponentially decreasing weight for past observations to forecast in the future. In other words, the more recent the observation the higher the associated weight.
Model formulation: $\hat{Y}_{t+h\,|\,t} = l_t$

Where $l_t = \alpha Y_t + (1-\alpha) * l_{t-1}$

After observing the decomposition charts, we realized that our data has both seasonality and trend component and so went ahead with ETS which is essentially exponential smoothening and adjusts for error, trend and seasonal components.
Pros: Since it gives more weight to recent values, the modelling produces more accurate forecasts. And since it gives more significance to recent observations it is much easier to understand and interpret the underlying phenomenon.
Cons: It cannot handle trends as well as some other forecasting techniques do. It is best for forecast that are short-term and in the absence of any seasonal or cyclical patterns.
Modified methodology: We also used blocked bootstrapping to generate new time series which is like our observed series. The underlying principle is to decompose the time series into its trend, seasonal and irregular component and then shuffles the irregular component to get bootstrapped

remainder series. However, we need to be cognizant of the fact that there would be autocorrelation between the irregular component and simply cannot re-draw data points from the time-series. Instead in blocked bootstrapping contiguous sections of data is drawn at random and joined. Then we add the original trend and seasonal component back to these bootstrapped remainder components.

RESULTS

Clustering:

After applying the above-mentioned clustering algorithm, we were able to plot elbow plots and do bootstrap evaluation of clusters and cluster validation for internal measures and stability. The methods utilized were K-means, CLARA, PAM, Hierarchical, and Spectral Clustering. Each clustering method was evaluated using resampling with subset, jitter, boot, and jitterboot. These methods are where the data is resampled with subset, noise, and addition of noise to existing data points to get a better measure of existing clusters.

The bootstrap evaluation of clusters is one of the better methods to check the stability of the clusters formed for the data. It provides us with the Jaccard Coefficient and number of a cluster was dissolved to form a stable cluster. We achieved clusters with Jaccard coefficients more than 0.7 using the method PAM which also provided us with most meaningful clusters. Jaccard coefficient is explained in our literature review. We used clValid() package in R to validate the cluster in terms of Dunn index, Connectivity, Average Distance, and Silhouette width.

The table below provides us with the detail about the methods utilized here:

Clustering Methods	Maximum Jaccard Coefficient	Minimum Jaccard Coefficient	Dunn	Silhouette	Connectivity
PAM	0.867	0.7	0.0149	0.41	91.53
K-means	0.74	0.503	0.0602	0.4889	59.035
CLARA	0.88	0.49	0.0173	0.447	84.111
Hierarchical	0.93	0.505	0.0906	0.43	37.56
Spectral	0.745	0.63			

Table 3. Clustering Results

Optimal Number of stable clusters for PAM:

Number of clusters	Maximum Jaccard Coefficient	Minimum Jaccard Coefficient	Maximum Dissolution	Minimum Dissolution
5	0.76	0.508	8	63
7	0.78	0.505	60	10
10	0.867	0.7	37	9
12	0.84	0.56	53	7

Table 4. Optimum number of Clusters

Optimal Number of clusters by methodology:

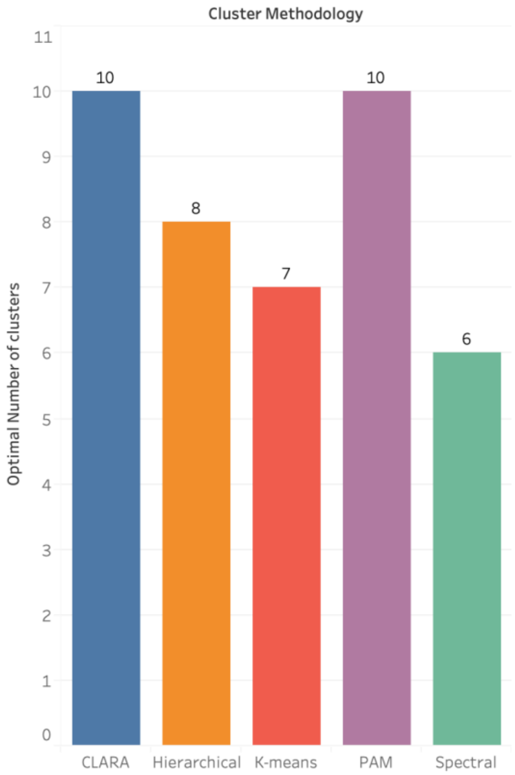


Fig 5. Optimal number of clusters by clustering methodologies

Store distribution across clusters:

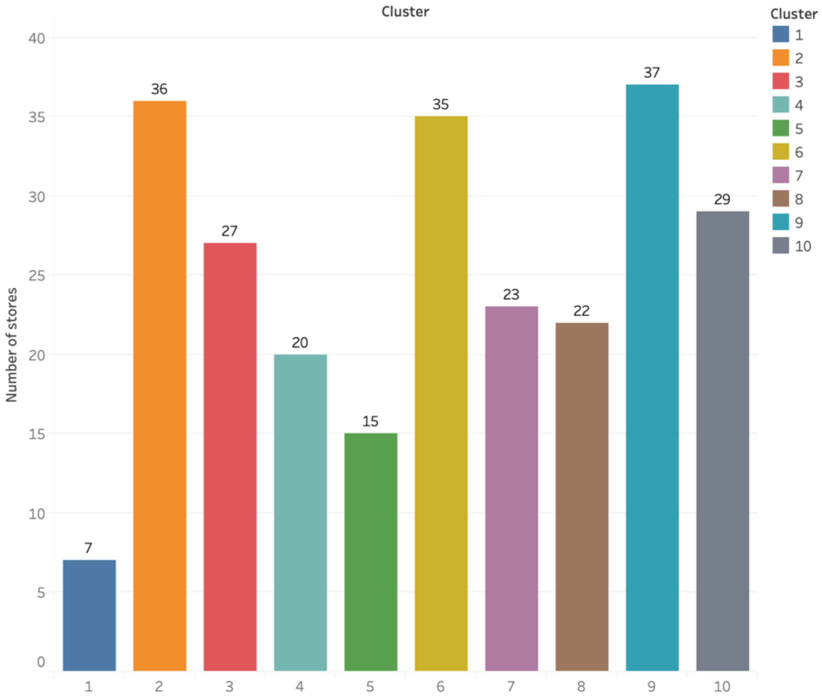


Fig 6. Store distribution across clusters

The geospatial mapping of the cluster:

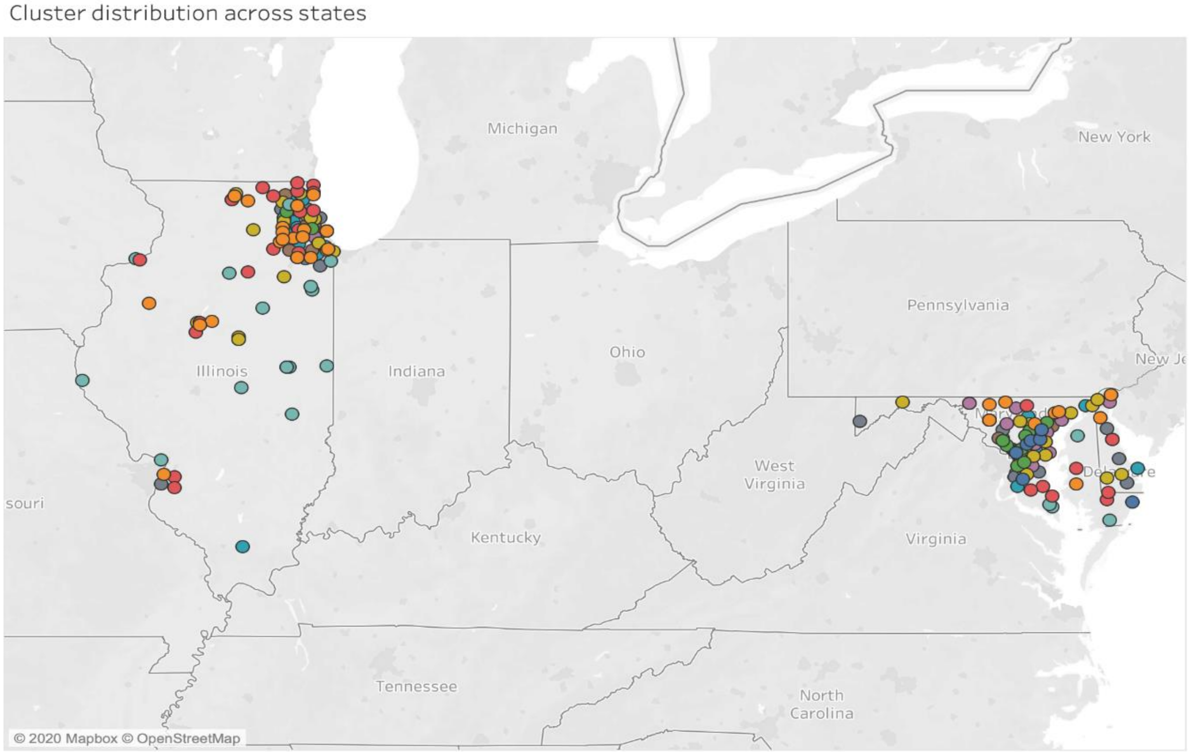


Fig 6. Geospatial distribution of stores

Time series:

We perform univariate time series forecasting using multiple advanced modelling techniques including ARIMA, ETS and HW. We also bootstrap the time-series to further bolster our results. A comparison between bagging of forecasts versus without bagging of forecasting is summarized in the following sections.

1. **Bagging Vs Without Bagging:**

a. **Fig 7:** Bootstrap example for Cluster 6, ten generated time-series through bootstrap identical to our observed time series

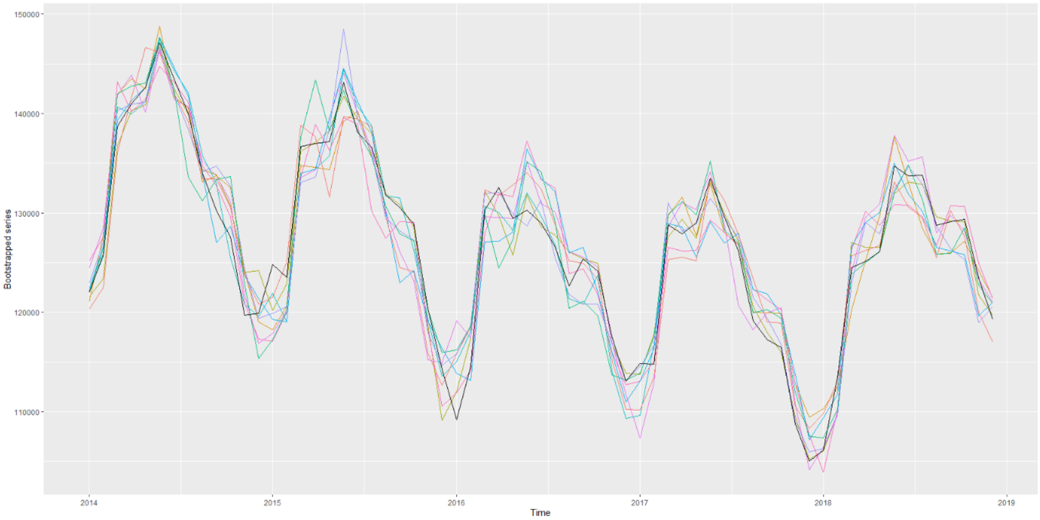


Fig 7. Bootstrap data for Cluster 6

b. **Fig 8:** Bagging example for cluster 6, where the forecast of all bootstrapped simulation is bagged to generate an aggregated mean forecast

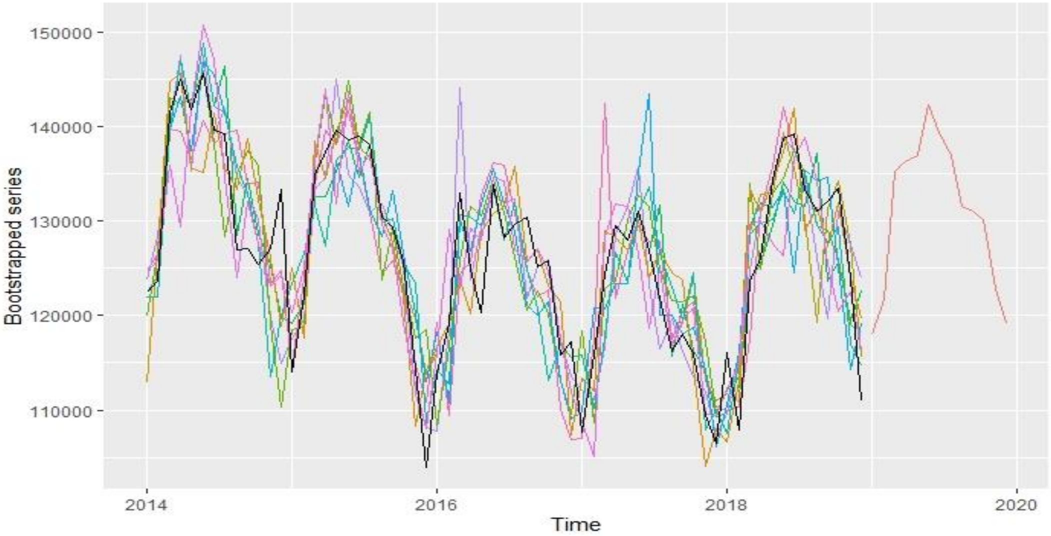


Fig 8. Bagged data for Cluster

c. **Fig 9:** Visual Forecast accuracy comparison between Bagged ETS forecast Vs without bagging ETS forecast. As you can observe in the graph below that Bagged ETS forecasts are much better than normal ETS forecasts

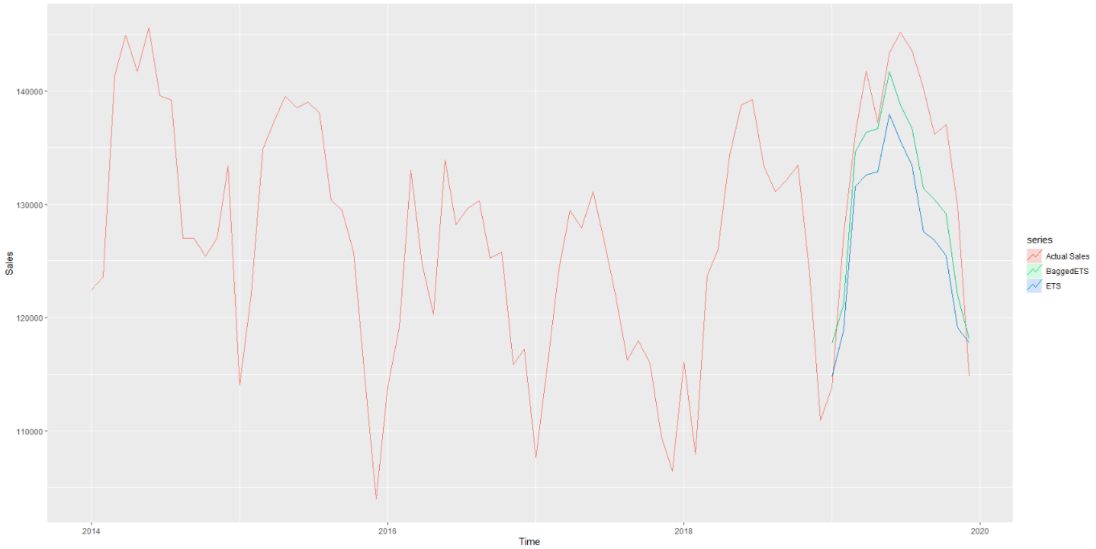


Fig 9. Forecast of Bagged ETS vs ETS

2. **Accuracy analysis**
   For the univariate time series forecasting we used Mean Absolute Percentage Error as our statistical performance measure to compare models. Comparison are displayed below in Table 1. Store 7521 and Store 7848 which opened in 2017 and tagged as new stores have the best accuracy in the test period. However, Store 7755 and Store 7910 have poor accuracy in the test period.

a. **Table 1:** MAPE for new stores across all forecasting techniques for the test period (first 13 periods of 2018) and train period(first period of 2014 through corresponding period of 2017)

Model	Period	Store 7521	Store 7848	Store 7755	Store 7910
HW	**Test**	90.4%	87.6%	38.6%	0%
Bagged HW		90.5%	88.8%	39.5%	0%
ETS		90.4%	91.9%	28.7%	0%
Bagged ETS		90.6%	91.3%	35.0%	0%
ARIMA		91.7%	90.3%	42.4%	0%
Bagged ARIMA		91.2%	89.9%	40.4%	0%
HW	**Train**	97.4%	97.3%	97.5%	97.3%
Bagged HW		97.6%	97.5%	97.7%	97.5%
ETS		97.5%	97.3%	97.2%	97.3%
Bagged ETS		97.5%	97.6%	97.8%	97.6%
ARIMA		97.9%	97.4%	98.0%	97.4%
Bagged ARIMA		97.5%	97.4%	97.5%	97.4%

Table 5. Train vs Test Results of Forecast Models

3. **New Store Analysis:**

   a. **Store 7521:** As you can observe below the Store and Cluster mean sales are in-line.

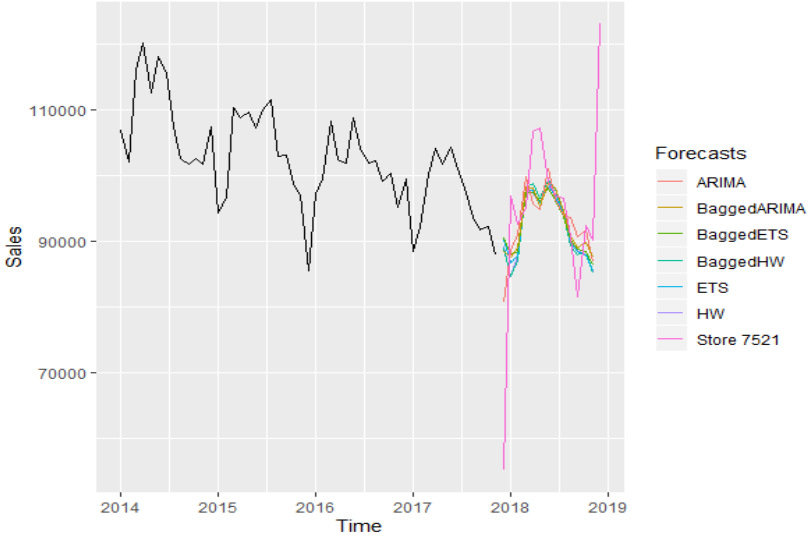


Fig 10. Actual vs Forecast for store 7521

   b. **Store 7848:** As you can observe below the Store and Cluster mean sales are in-line.

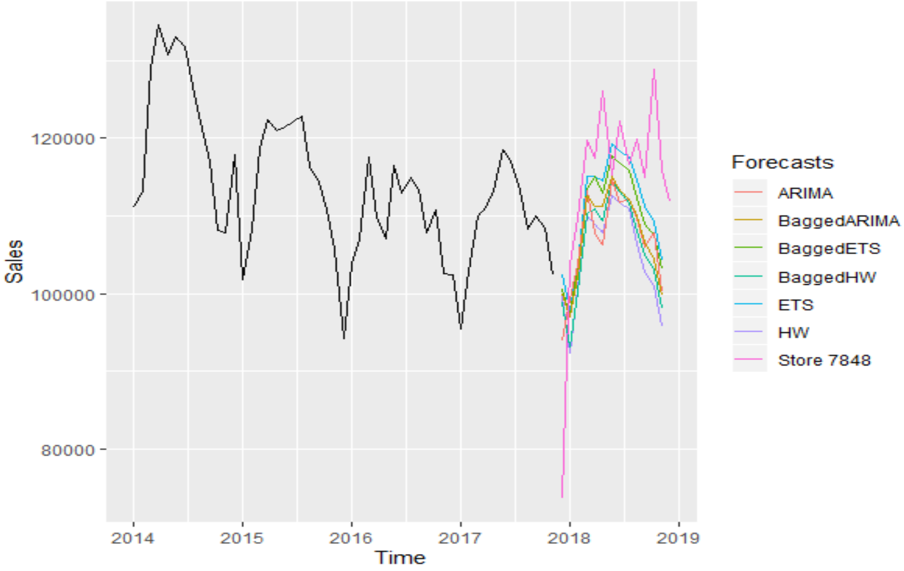


Fig 11. Actual vs Forecast for store 7848

c. **Store 7755:** As you can observe below the Store and Cluster mean sales are NOT in-line.

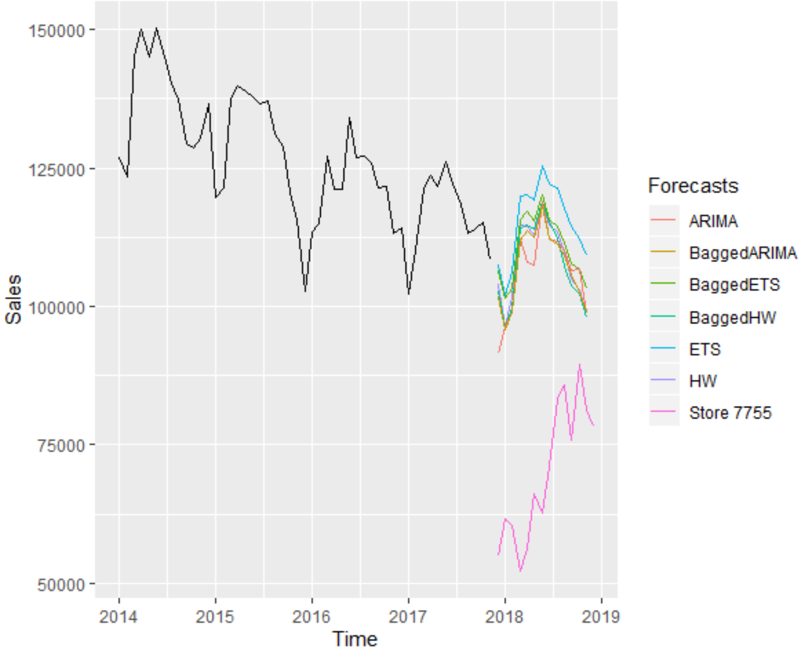


Fig 12. Actual vs Forecast for store 7755

d. **Store 7910:** As you can observe below the Store and Cluster mean sales are NOT in-line.

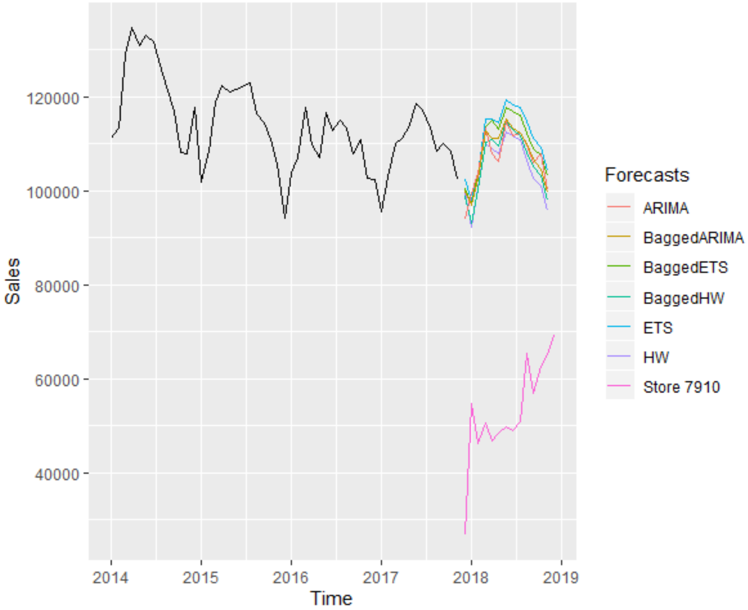


Fig 13. Actual vs Forecast for store 7910

CONCLUSIONS

Huge investments are committed when organizations make decisions on opening a new store and return on investment becomes a crucial factor while choosing a new site. Accurate sales forecasting for a potential new site thus becomes essential for a new location planning decision. Data-driven analytics can leverage existing data to enable clustering of existing stores and forecasting for new stores considering the various demographics factors in play.

Our solution adjusts for multiple demographic variables and successfully clusters existing stores which then can be leveraged to deliver substantially better forecast. Forecast accuracy improves by ~9 percentage points using clustering versus forecasting without clustering or using other heuristic approaches.

Our solution works well with forecasting the sales of two of the tagged new stores 7521 and 7848, however, the solution did not deliver high accuracy for tagged new stores 7755 and 7910 where the forecast accuracy is below 50% for both. We have assumed the mean sales of a cluster as the past historical data for our tagged new stores. Since a cluster has more than 20 stores on average with each of these stores having varying sales across the years, computing an aggregated sales value for each cluster is not optimized and needs further deep dive. There is scope of improvement in the clustering of all stores and how the aggregation of sales is computed for each cluster.

REFERENCES

Binalabid, F. (2014). A Novel Approach for PAM Clustering Method. *International Journal of Computer Applications*, *86*(17), 1–5. doi: 10.5120/15074-3039

Hülsmann, M., Borscheid, D., Friedrich, C. M., & Reith, D. (2011). General Sales Forecast Models for Automobile Markets Based on Time Series Analysis and Data Mining Techniques. *Advances in Data Mining. Applications and Theoretical Aspects Lecture Notes in Computer Science*, 255–269. doi: 10.1007/978-3-642-23184-1_20

Na, L., Ying, C., & De, B. (2019). The empirical analysis of convenience store sales forecast based on multidimensional data in emerging markets. *Proceedings of the 3rd International Conference on Business and Information Management - ICBIM 19*, 1–5. doi: 10.1145/3361785.3361802

Rogers, D. S., & Green, H. L. (1979). A New Perspective on Forecasting Store Sales: Applying Statistical Models and Techniques in the Analog Approach. *Geographical Review*, *69*(4), 449–458. doi: 10.2307/214807

George, L. L. (2004). Actuarial Forecasts for the Automotive Aftermarket. *SAE Technical Paper Series*, *113*(5), 697–701. doi: 10.4271/2004-01-1538

Wood, S., & Tasker, A. (2008). The importance of context in-store forecasting: The site visit in retail location decision-making. *Journal of Targeting, Measurement and Analysis for Marketing*, *16*(2), 139–155. doi: 10.1057/jt.2008.3

Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*. doi: 10.1016/j.ijforecast.2019.06.004

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271. doi: 10.1016/j.csda.2006.11.025

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster validity methods. *ACM SIGMOD Record*, *31*(2), 40. doi: 10.1145/565117.565124

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting principles and practice*. Melbourne: O Texts, online open-access textbooks.