Forecasting Claims for Extended Warranties

# Final Report - BAN 620: Data Mining

Bhavya Raj Bhati

[**Final Report - BAN 620: Data Mining**](#_lbsw91ewckuh) **1**

[Introduction:](#_jy35w1tum04w) 2

[Data Pre-Processing](#_r6hl4xtv1jhc) 3

[Forecasting Methods](#_yms27xxusaoe) 4

[MVR:](#_e7qcrzyztwoq) 4

[k-Nearest Neighbors:](#_85kylobbwvl) 4

[Time series regression:](#_e0iwklk3cjmm) 4

[K-NN with Time Series Regression:](#_vt3o9i5g5evd) 4

[Accuracy Metrics for Comparison:](#_7ubfig9215af) 4

[MultiVariate Regression](#_600pdk5h06e5) 5

[K-NN with Time Series](#_yfokbl4mm8cm) 5

[Creating the Model:](#_y39f0ns66wyg) 5

[Our Model:](#_epumua9r61d) 5

[Predictions for Future Months](#_1k8o21aztcf6) 6

[Backtesting:](#_wt2ojsxpzx0x) 6

[Our Predictions:](#_bjt5kbc35s5a) 7

[6 months predictions:](#_nkfafqv37243) 7

[12 months predictions:](#_4u65nt98liq6) 7

[Conclusion](#_p5liqa9f5tn3) 8

## **Introduction:**

A customer purchasing items at a store is often given the opportunity to purchase an extended Warranty at the time of purchase. If the item malfunctions (or breaks), depending on the coverage of the warranty, the customer can file a claim with the extended warranty company to have the item repaired or replaced or get a payout. This makes forecasting claims extremely crucial for the extended warranty company. We make an attempt to forecast claims using a company’s historical data on claims, sales and active warranties. For this purpose we choose MultiVariate Regression and Time Series Regression with K-NN, which are data mining techniques to see which turns out better. We are using data from the past 3 years from the start of 2017 to the current data and aggregated it to monthly. and we are trying to predict the next 6 months. and therefore we are choosing validation data for 6 months that is roughly 10% of the dataset. To evaluate our predictions we chose to look at Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Finally we predict the results for the next six months and the next twelve months. As a check for our accuracy of the model we backtest it and evaluate the model based on the same accuracy metrics mentioned above.

It is significantly important to note that there is an order of magnitude difference between the variables which means that we will need to scale and normalize our variables. All computations and model building will be done in R.

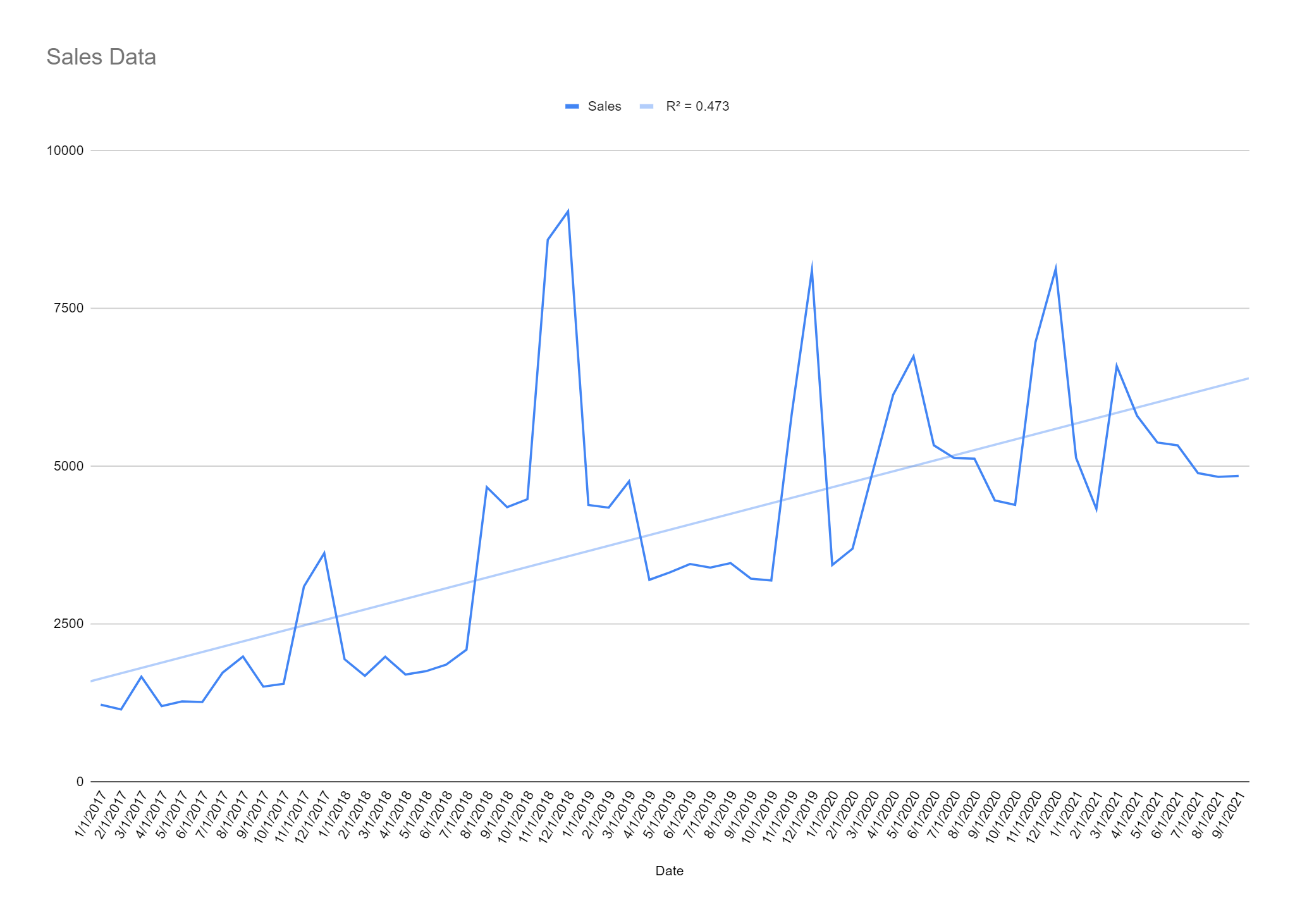
## **Data Pre-Processing**

The data was collected from the database of an extended warranty company. Due to various restrictions of not being able to share the whole dataset for this project we were unable to get all the variables such as warranty term or product mix or retailer mix. The company did provide aggregated data on their warranties and claims. The data collected through a series of SQL queries which meant that we didn’t have to deal with anomalies or missing values. Since SQL is out of the scope of this project, we won’t share the scripts.

To explore the data further we were able to make plots and see the trends. The charts are shown below and we can clearly see that although a lot of the data on the claims side can be explained by a simple trendline.

|  |  |
| --- | --- |
| Chart | Chart |

As we explore the data we were able to determine that an MVR or a time series model might work best. Below we can see the plot for Sales as well.



## **Forecasting Methods**

### MVR:

When there is more than one predictor variable in a multivariate regression model it will be useful. As we are predicting claims from Sales and active warranties, we tried using MVR model here and comparing accuracy of model with other models.This method involves creating a regression model between Claims and the other variables.

### k-Nearest Neighbors:

KNN regression is a **non-parametric method** that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. We treat Claims as a time series and run a K-NN regression.

### Time series regression:

Time series regression is **a statistical method for predicting a future response based on** the response history (known as autoregressive dynamics) and the transfer of dynamics from relevant predictors. We are predicting Future claims and based on past data of sales and active warranties so Time series would be best fit for this project.

### K-NN with Time Series Regression:

We will be using the tsfknn function in R to make a model that incorporates K-NN while doing a time series analysis. The reason we do this is because from our data exploration we know that a time series would work and K-NN uses nearest neighbors in it’s regression so potentially a model incorporating should yield better results.

### Accuracy Metrics for Comparison:

Accuracy: For comparisons between our models we have chosen to go with Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). These are the performance KPIs that will be used to decide our final model, especially MAPE.

## **MultiVariate Regression**

* Fundamentally speaking Active Warranties would scale with Sales with respect to the mix of the terms of respective warranties sold.
* Hence in model Y variable would be Claims and X variable would be Sales and Active warranties.
* MODEL: MVR = **lm(Claims ~ Sales + Active.Warranties, data=train.data)**
* We are Predicting claims for next 6 months and we have about 59 records so we are taking 10% of data as validation data which is roughly 6 months.
* To check accuracy are comparing valid data and data which we got from a model based on train data using the **predict** method(R code).And then checked accuracy between predicted Claims and actual Claims.
* We are getting following accuracy method,



## **K-NN with Time Series**

### Creating the Model:

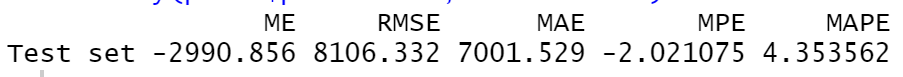
To create our model we partitioned our data into test and validation data. Our total number of records were 58. In a time series model it is very important to train your model the best you can with the most amount of data available to have the best results. We had the first 52 records for our training data. Thus we decided to use 6 months of validation data. This is not ideal but a trade-off to have more training data. In this instance it should be fine because we intend to start off with only predicting the next six months.

### Our Model:

We use the *tsfknn()* function in R programming language to create our model. Here is the final model statement we came up with:

knn\_forecasting(ts(train2.data), h = 6, lags = 1:6, k = 10, transform = "additive")

This simply means that we are using 6 ‘lags’ or lagging points to create a time series relationship to come up with a model. The best accuracy we get for this model is when k=10. From the claims curve we can observe that the trend is not exponential but is increasing which means that the transformation in our function should be additive. h = 6, simply means that the prediction using this function will be for the next six months. When we test our model’s predictions against the validation data for six months we get accuracy metrics as:



As we can see from the above accuracy metrics, this model has a lower RMSE and MAPE than our MVR model. This is the model that we should go with to predict claims for the company.

### Predictions for Future Months

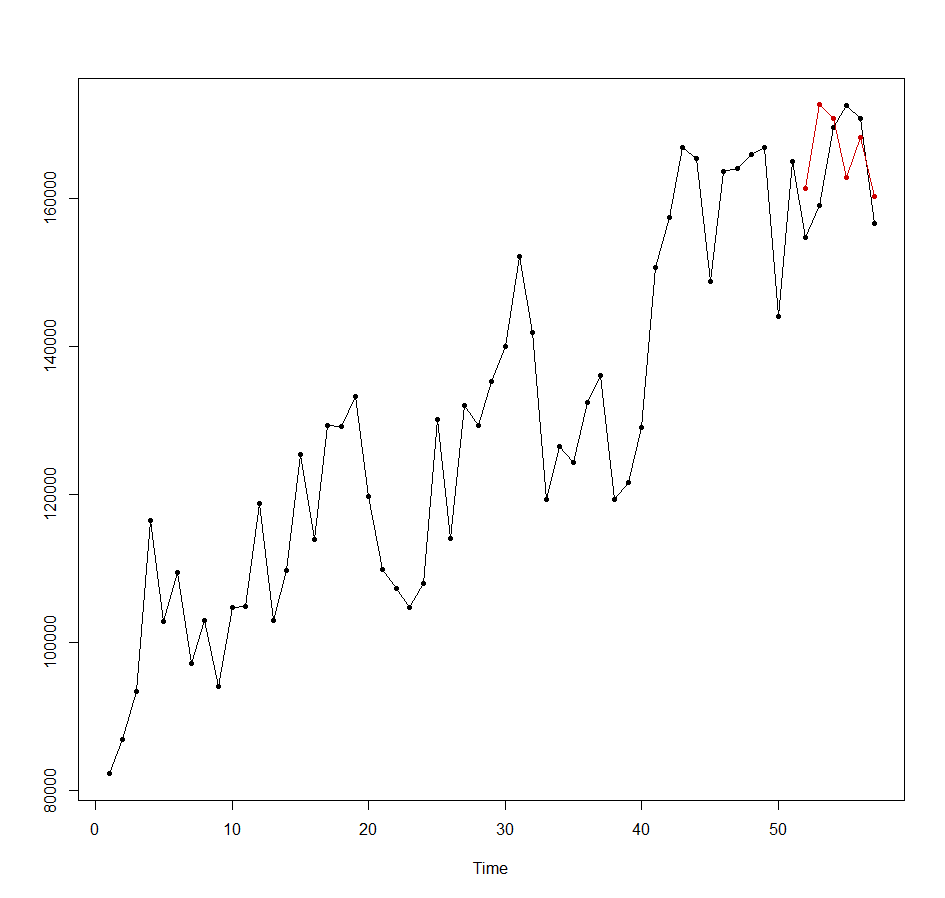
In order to predict the next six months using K-NN with a time series model we have to run our model again on the entire dataset. We do this using the following model statement:

knn\_forecasting(ts(final.df), h = 6, lags = 1:6, k = 4, transform = "additive")

We use k=4 as this gives us the best accuracy on the entire dataset.

### Backtesting:

We used the entire dataset to create our new model so don’t have any validation data to test this on. However, we can backtest our model and see how it would have performed had we used it six months ago. We do this using the rolling origins function in R. When we do this we get the following plot and accuracy metrics:



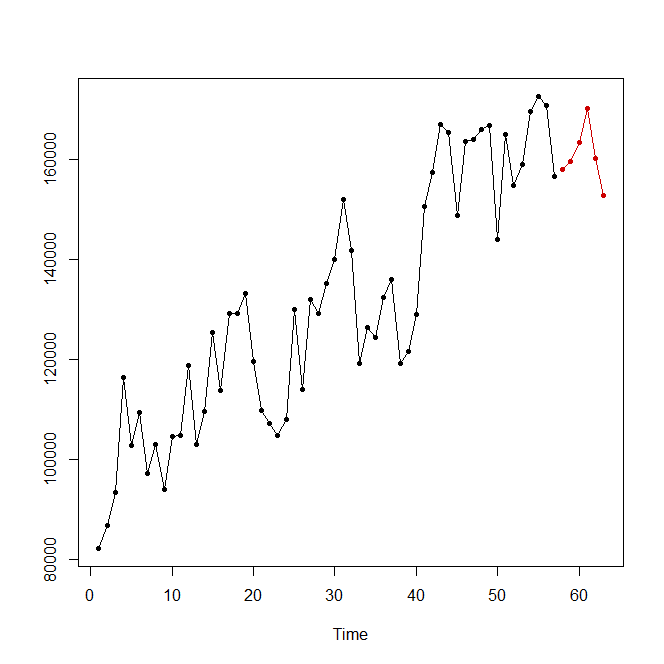


The red line represents the back tested data against the black line which is the actual claims data. The accuracy metrics show that we are 3.6% on MAPE and 6997 on RMSE. This means that this model is extremely accurate in predicting claims.

### Our Predictions:

#### 6 months predictions:

Below is a plot of the next six months predictions using our new model:



1.October 2021: 157,969

2. November 2021: 159,571

3. December 2021: 163,417

4. January 2022: 170,199

5. February 2022: 160,107

6. March 2022: 152,813

#### 12 months predictions:

The plot on the left shows the backtested values against the actual claims values. The backtesting was done using the rolling origins function. The accuracy metrics above show that while our current model is still very good at predicting claims twelve months out it performs better at shorter time frame predictions.

## **Conclusion**

The dataset we picked was to predict claims of an actual warranty company from the industry to gain experience in a real world problem and how to perform an analysis using data mining techniques to answer real world business problems. What we can learn from this analysis is that the company’s claims can be very accurately forecasted over six month forecasting periods. We set out to forecast the company’s claims over the next six months. We tested MVR and K-NN with Time Series models. From our tests we were able to conclude that K-NN with Time Series model is the best one to use.

We present our predictions for the next six months after running our model on the entire dataset. Our accuracy metrics when backtested give a very good MAPE of 3.61% and RMSE of only 6997. While our accuracy metrics from a twelve months predictive model were promising we would recommend using our models for six month predictions. Finally we can confidently conclude from the data that, with the dataset we were given, we can predict the company’s claims for the next six months with a high level of accuracy.