For this final project we have tried various models. We started by considering Random Forest which failed for us by our approach and then we continued exploring more models like Regression Models, Decision Tree Classifier using GridSearchCV Model but these too failed for us as though it worked well for few steps. Then we started working on Decision Tree Classifier native model and that worked well for us. We tried data exploration in R too but we stuck at some point of time so we moved to python and we tried implementing it in python moving ahead.

Below were the models we tried and out of which Decision Tree Classifier model worked best for us.

We started the project with Data exploration exercise using R (Data\_Exploration\_In\_R.ipynb) in which we grouped it under (shop, item) totals and finding frequencies.

Major finds are as follows :

(1) 30% of the frequencies (of unique shop, item pairs) were of "1" count sold. And 65% were between "1" and "6" counts sold. This distribution showed a very "poisson" like histogram.

(2) Shapiro "normalization" test for total item counts for every pair (shopID, ItemID) showed "Reject H0 for normalized test"

Then we started implementing our models in python and we started with Random Forest (please see : FAILED\_Random\_Forest.ipynb) where due to high root-mean-square-error we had to reject the model.

Then we moved towards Regression models (please see : FAILED\_Regression\_Models.ipynb).

For all the above failed models we ran “evaluation” only on "training" data. They were rejected so never tried on test data.

Generalized regression models and Random Forest Regression model returned very low R-squared value (Coefficient of determination) and a high root-mean-square value. Thus these models were rejected. Here’s what you can see in ipynb files of the “FAILED” models.

(1) Extracted features from training data like the (ShopID, ItemID) to predict monthly totals.

(2) Optionally normalized/Standardized the "features vector"

(3) Tried without success following models in that order, all showing low R-squared (close to zero).

(a) Linear Model

(b) GLM with various distribution and link function combination: (gaussian, identity), (gamma, identity), (Poisson, log) all without success.

(c) Random Forest Regression model: This model resulted in root-mean-square value of 3.4, which is high.

Following regression models, we moved to Decision Tree Classifier using GridSearchCV model. (FAILED\_Predict\_Future\_Sales\_Using\_Decision\_Tree\_Classification\_GridSearchCV\_Model.ipynb)

When we observed the steps, it looks work well with beats score of around 0.65 for max\_depth=3 but when we check the predicted result, i.e. for all item\_id for future prediction date\_block\_num=34, the count was 1 which put this model in doubt for us.

Then we moved to native Decision Tree Classifier model which worked really well for us though accuracy is 0.54 but the model predicted really well and were able to get the predictions as saved in Prediction\_using\_decision\_tree\_classifier.csv.

Predicts can be seen in below plot:

A picture containing text, businesscard

Description automatically generated

We tried to identify that how many items as per item would be sold in future month irrespective of shop\_id and below was the plot:

Scatter chart

Description automatically generated with medium confidence