**Team Description**

Team Name – The Learning Machines

Team Members - Ang Zhen Xuan, Phang Chun Rong, Tan Qing Yang

**Description of Work allocation**

1. Zhen Xuan – Provision of reading resoures, did comparison of models’ performances through restructuring templated cross validation code. Writeup of the overall workflow in the README.
2. Chun Rong – Templated cross validation code, refactored code structure and its’ readability
3. Qing Yang – Explored into modelling using XGboost and did testing of models

**Machine Learning - Rossman Store Sales**

**Exploratory Data Analysis & Feature Engineering**

* Label encoded features with classes/categories & filled NAs with 0.
* Removed stores which are not opened (and have zero *Sales*) in our training set as we realized they do not help much in our model’s training phase.
* Appended the store.csv features to the respective *Store* column in train.csv
* Constructed additional features through engineering of existing features such as *WeekOfYear*, *Day*, *Month*, *Year,* etc.

**Modeling Phase**

**IF sufficient time, we can have our own holding test set, by breaking our train data into train, test (holdout set so we don’t overfit to all our training examples)**

We adopted the k-fold cross-validation procedure in our model training to pick the best model to be used in our final evaluation for the Homework’s holding out test set.

Since we are dealing with time-series data, the conventional k-fold cross-validation approach does not account for the ‘time’ aspect of our data. The methodology we have followed is as such:

enter image description here

This accounts for the ordered series of data we are dealing with.

In addition, the predictions we plan to obtain are of continuous values (not categorical), so the algorithms we adopted solves the regression issue. These algorithms include the *Support Vector Regression*, *Random Forests Regressor* (from the *scikit-learn* libraries) as well as the *Gradient Boosted Trees* (from the *xgboost* libraries) algorithms.

Our k-fold cross-validation approach uses the RMSPE metric that our model is being judged on to pick the best model, as well as the optimal values for the tunable parameters for each model. We run the k-fold cross-validation for 20 times to get a stabilized average RMSPE value for us to properly judge the optimal model (in terms of tunable parameters) on. **BEST MODEL IS XXXXX, we use that for leaderboard testing.**

**IF ENSEMBLE OF OUR THREE MODELS IS GOOD ENOUGH WE CAN USE IT AS FINAL MODEL THEN EXPAND ON EXPLAINING HERE.**

**Statement of Individual Work**

Please initial (between the square brackets) one of the following statements.

[AZX-PCR-TQY] We, A0139569L-A-A0139Y, certify that I have followed the CS 3244 Machine Learning class guidelines for homework assignments. In particular, I expressly vow that I have followed the Facebook rule in discussing with others in doing the assignment and did not take notes (digital or printed) from the discussions.

[ ] We, A0139569L—A0139Y, did not follow the class rules regarding the homework assignment, because of the following reason:

<\*Please fill in\*>

I suggest that I should be graded as follows:

<\*Please fill in\*>

**References**

I have referred to the following list of people and websites in preparing my homework submission:

Useful references from past winners

* <http://blog.kaggle.com/2015/12/21/rossmann-store-sales-winners-interview-1st-place-gert/> - first place overview
* <http://mabrek.github.io/blog/kaggle-forecasting/> - top 10%
* <http://www.elasticmining.com/post/2016-01-02/time-series-forecasting-kaggle.html> - simple intro to time series modelling
* <http://blog.kaggle.com/2016/02/03/rossmann-store-sales-winners-interview-2nd-place-nima-shahbazi/>

Feature Engineering

* <https://www.kaggle.com/c/rossmann-store-sales/discussion/17048>

Model Selection

* Why no need for linear kernel -  <https://stats.stackexchange.com/questions/73032/linear-kernel-and-non-linear-kernel-for-support-vector-machine>

Cross Validation

* <https://stats.stackexchange.com/questions/14099/using-k-fold-cross-validation-for-time-series-model-selection>
* <https://stats.stackexchange.com/questions/82546/how-many-times-should-we-repeat-a-k-fold-cv>

Ensembling methodologies

* <https://mlwave.com/kaggle-ensembling-guide/>
* <https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation>
* <https://stats.stackexchange.com/questions/2306/feature-selection-for-final-model-when-performing-cross-validation-in-machine?rq=1>