Machine Learning Assignment 1

Manas Gaur, Ph.D student, Computer Science, Wright State University

September 22, 2017

1 Introduction

Linear Regression In this we will be exploring different types of regression functions in machine learning. Such as , linear regression with one variables, quadratic regression and Linear regression with two variables. Our analysis and modeling will be using flu dataset. The flow of the assignment is organized as follows; Firstly we will be providing a small descriptive analysis of the dataset. Secondly, modeling and analysis of linear regression with one variable. Thirdly, modeling and analysis of quadratic regression and Finally, modeling and analysis of linear regression with two variables. The dataset provided has 417 data points and 18 features (we consider Risk as the target variable in this assignment). In 2011, using the self-report survey of 410 high school students in the mid-west, the data was collected. Analysis of the dataset showed that data has 56% of missing values. We calculated the percentage of missing values by summing over the data points containing *no-values* and *NaNs*. Our experiment in this assessments focus on impact of flu transmission knowledge (*KnowlTrans*) and Respiratory etiquette (*RespEtiq*) on perceived risk of contracting influenza (*Risk*). We enlisted the descriptive values of the features in the table 1.

Table 1: Features and Data type

Feature Names	**Data Type**	**Type of Variable**	**Range**
KnowlTrans	Logits	Independent	[-1.393,1.393]
RespEtiq	Logits	Independent	[-1.453,1.453]
Risk	Likert Scale	Target	[1 (low),2,3,4,5 (high)]

In this assignment, for making the linear regression learn we are using stochastic gradient decent and cross validation.

1.1 Stochastic Gradient Descent

This gradient descent is used when we have small training examples. The reasons of using stochastic gradient descent (SGD) are: reducing the variance while updating the parameter (θ) and leading to stable convergence.

$$\theta = \theta - \alpha \nabla_\theta J(\theta; KnowlTrans^i, Risk^i)$$

where, $(KnowlTrans^i, Risk^i)$ are from the training set. α is the learning rate and θ are the parameters of the features (tells which features is really impactful/important for predicting Risk). Learning rate (α) is one of the sensitive parameter in the SGD. It affects the convergence rate. A high value of learning rate leads to early convergence and poor training while very low learning rate delays the convergence. In this assessment we will observe the impact of learning rate on the model learing in various scenarios. We organize these scenarios as follows : section 2, will experiment with linear regression using one variable, section 3 will experiment with linear regression with quadratic features and section 4 provide a tabular analysis of the model when we split the data in various ratios and finally, we provide cross validation results and compare the model prediction in 3 scenarios by plotting the model predictions over observed values. **Note : We haven't performed outlier detection in this assessment**.

2 Linear Regression with One Variable

In this section, we will experiment linear regression using the flu dataset, considerable only one independent variable *KnowlTrans*. This feature provides knowledge about, how the flu is transmitted. The values in this feature are score in logits. The target variable is *Risk*. Target variable defines perceived risk of contracting influenza. The values in this variable is measured in logits. One thing that is evident, both the independent feature (*KnowlTrans*) and target variable (*Risk*) have same type of data-values. Let's visualize the behavior of these two variable using the notion of cumulative sum and co-variance.

In the figure 1, we observe that *KnowlTrans* and *Risk* are negatively correlated over the entire observations. The calculated covariance between *KnowlTrans* and *Risk* is **-10.8978**. We will be analyzing the performance of linear regression using *root means square error* (RMSE). RMSE is defined as standard deviation of residuals (prediction error). It is defined at the distance between the data points and regression line.

$$\text{RMSE} = \sqrt{(Predicted - Observed)^2}$$

We considered two approaches in assessing the performance of linear regression model.

- Linear Regression without Normalization
- Linear Regression with Normalization using Min-Max
- (Not sure whether I will be able to do it in time) Linear Regression with Normalization and Regularized Stochastic Gradient Descent.

In the approach of modeling without normalization, we considered the data points as present in the dataset. In the section 2,3 and 4 we have used normalized data for model training and testing.

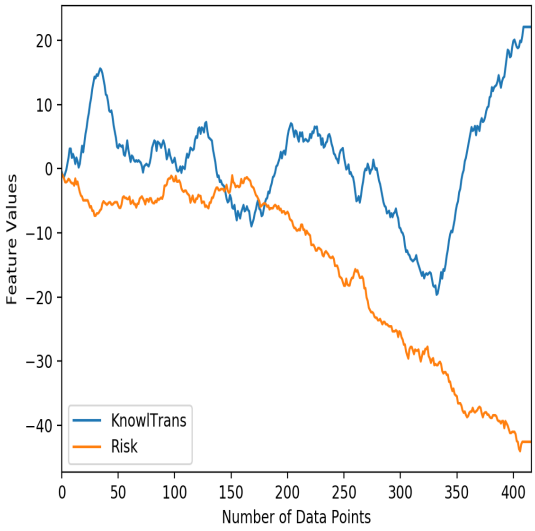


Figure 1: Individual cumulative sum of KnowlTrans and Risk over all the data points. Data points means observations.

2.1 Linear Regression without Normalization

The values in *KnowlTrans* lies in the range [-1.393,1.393] and values in *Risk* lies in the range [-1.453,1.453]. Certainly, seeing these ranges, one can suggest the approach of Normalization. Cornerstone of our methodology is to see, how much difference occur in the error , when linear regression is trained over normalized and non-normalized data. This is because, normalization process is a time consuming process.

In the figure 2, we are analyzing the prediction behavior of the model by varying the learning rate. We observed that , at very high learning rate, the prediction error shoots up very high. In order to prevent such abnormal bump, we need to keep the learning rate optimal. Such behavior is attributed to poor training as shown in figure 3. The straight line in figure 3, shows that model failed to learn from training dataset. A general trend of asymptotic decline with flattening, proves sufficient training by the model. We trained the model with low learning rate of 0.01, and increment it with step size of 0.01 after 10000 epochs. We observed that impact of lowering the learning rate on the prediction error (y-axis) in the figure 4. When we reduced the learning rate for the training the model, we saw the near hard asymptotic behavior in the cost function. As shown in figure 5, there is sharp decline in the cost function value and flattening. From figure 4 it is observable that with increasing the learning rate of the model, there is increase in the prediction error. The zig-zag trend in the graph is attributed to changing learning rate , changing epochs while training the model and taking the average error (RMSE) over the training. We further

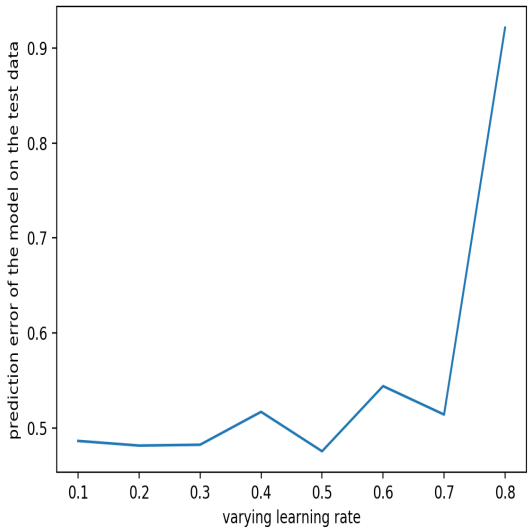


Figure 2: Trend in prediction error of the model with varying learning rate and epochs. Learning rate is kept between [0.1, 0.9] and incrementing with step size of 0.1.Epochs were kept in the range [1000, 10000] with increments of 1000.

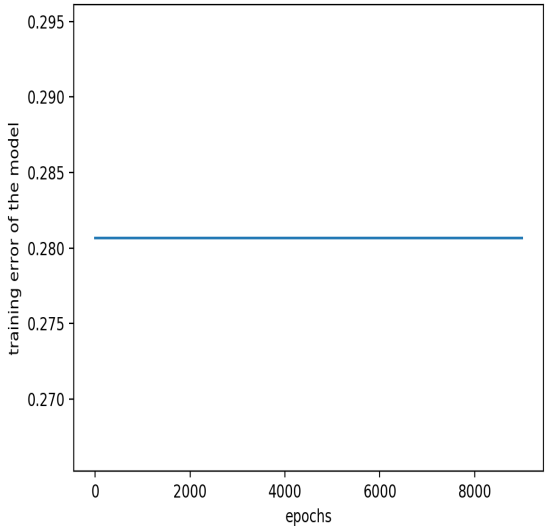


Figure 3: Training error of the model on the training set. The flat line shows high learning rate and poor learning by model. This training was performed over an un-normalized dataset.

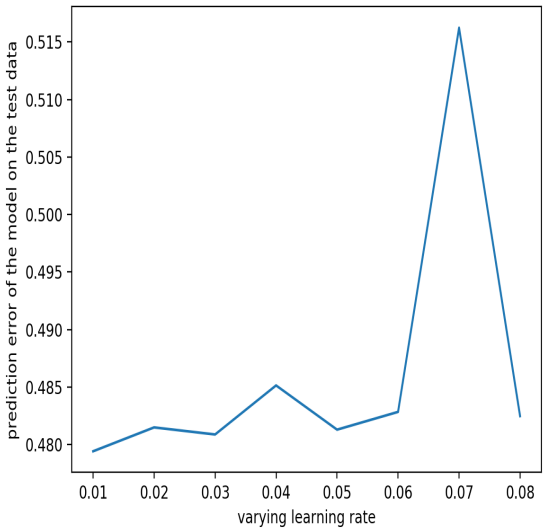


Figure 4: Trend in prediction error over changing learning rate and epochs. Learning rate varies in the range [0.01, 0.09] with step size 0.01. Epochs were kept varying in the range [1000 10000] and incrementing by 1000 .

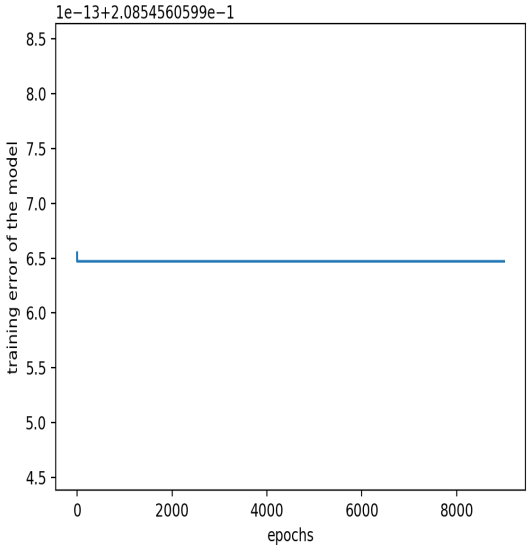


Figure 5: Training error of the model on the training set. There is small decline in the cost function (used to train the model) value before flattening.

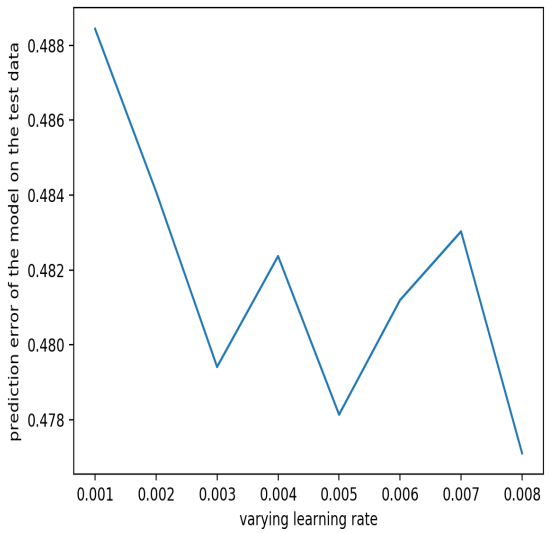


Figure 6: Trend in prediction error over changing learning rate. Learning rate varies in the range [0.001, 0.009] with step size 0.001.

reduced the learning rate of the model to 0.001 and increment it with step size of 0.001 after 10000 epochs. We observed similar behavior in the prediction error as in figure 4, as shown in figure 6, but there was significant change in the training error (y-axis). The reduction in the training error In the figure 7, the cost function does not show sudden convergence of the cost function. This behavior, claims that model got trained over the training dataset and it converged to an error value (value of cost function) of -0.750.
In our previous experiments, we tuned the parameters of the linear regression models and analyzed it's behavior on the dataset. We found that learning rate places a significant role in the training process of the model. Hence we carry out our future experiment with low learning rate. We did not normalized the dataset before training and testing the model over it. In the next experiment, we will train and test the linear regression model over the normalized data.

2.2 Linear Regression with Normalization using Min-Max

In this section, we will normalize the data (KnowlTrans, Risk) using Min-Max Scaling. This is a feature scaling procedure using in the pre-processing stage of KDD [1] cycle. It scales and translates each individual feature in the dataset, so that the values of scaled features lies in the range [0,1].

$$\text{scaled-dataset} = \frac{feature^i_j - \mu_{feature^i}}{max(feature^i) - min(feature^i)}$$
$$\text{i } \epsilon \{KnowlTrans, Risk\} \text{ and j } \epsilon [0, \text{length}(feature^i)\text{-}1]$$

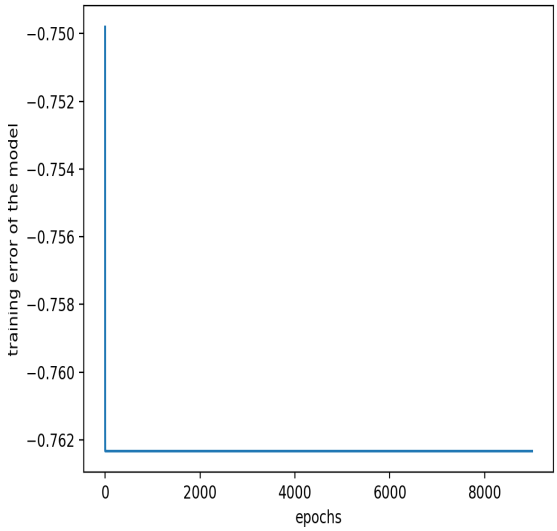


Figure 7: Training error of the model on the training set. There is steep decline in the cost function value before flattening, showing sufficient training before converging.

The normalization of the dataset affects prediction error, it is shown in figure 8, wherein the prediction error is close to 0.1. Normalization procedure, helps the algorithm identify patterns in the dataset so that it can fit the curve (regression line) correctly. Plot of the training error in figure 9, showed convergence with slight increase in training error but decrease in the prediction error. The increase trend in training error is attributed to Min-Max scaling which tend to alter the values of the features and bring them in the range of [0,1]. As stated before, the features in the dataset had values in logits ([-2,2]). **Note : Normalization of the dataset, does not affect the correlation between features and the target variable.**

3 Linear Regression with Quadratic Feature

In this section, we will be experiment the learning and predicting behavior of the model when another feature is added to the input which is a square of an existing feature. For instance; our new dataset for model contains features {KnowlTrans, $KnowlTrans^2$} with the target variable being same. So, we model *Risk* as :

$$\text{Risk} = \theta_0 + \theta_1 * KnowlTrans + \theta_2 * KnowlTrans^2$$

Even in, linear regression with quadratic feature, the covariance between features {KnowlTrans, $KnowlTrans^2$} and *Risk* is -11.8838 (negative correlation).

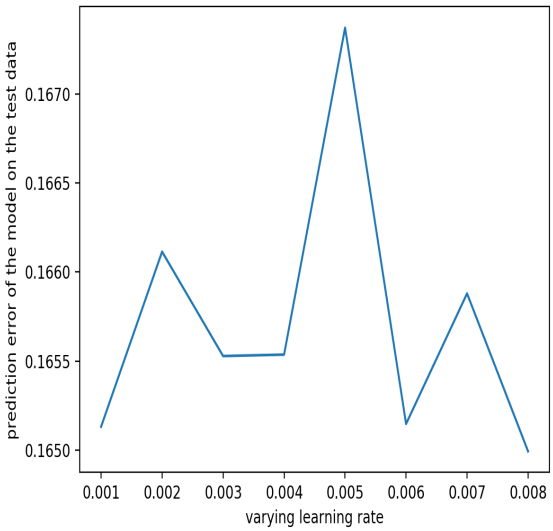


Figure 8: Trend in prediction error over changing learning rate. Learning rate varies in the range [0.001, 0.009] with step size 0.001. The prediction error range is between (0.15, 0.17).

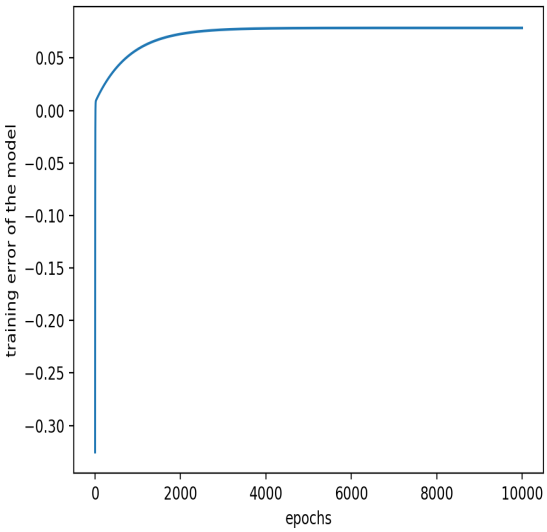


Figure 9: Training error of the model on the training set. There is steep increase in the cost function value before flattening, showing sufficient training before converging and the impact of bring the feature values from [-2,2] to [0,1].

Table 2: Contingency table for linear regression with Quadratic features in dataset

	KnowlTrans	KnowlTransSq	Risk
KnowlTrans	1.0	0.0914	-0.0672
$KnowlTrans^2$	0.0914	1.0	0.0051
Risk	-0.0672	0.0051	1.0

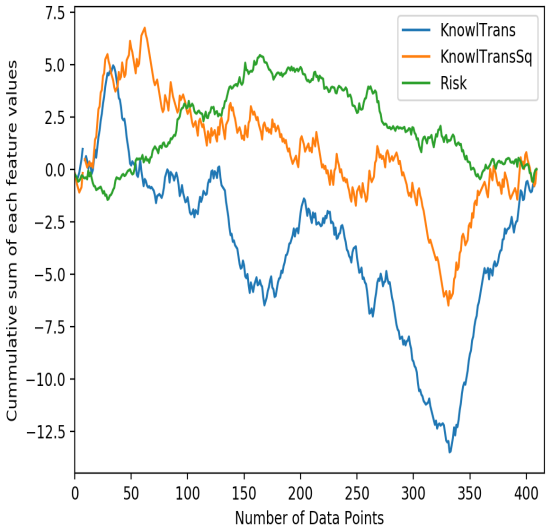
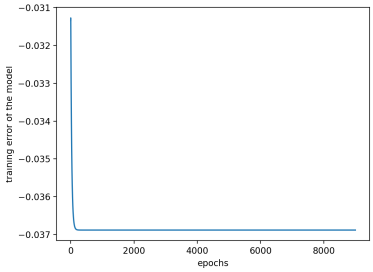
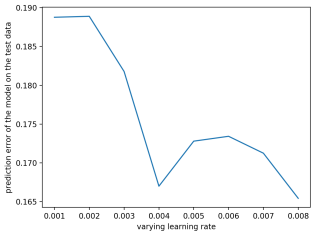


Figure 10: Individual cumulative sum of KnowlTrans, $KnowlTrans^2$ and Risk over all the data points.Data points means observations.



(a) Training Error



(b) Prediction Error

Figure 11: (a). Training error of the model when trained over a dataset with features (KnowlTrans, $KnowlTrans^2$). (b). Prediction error of the model.

In figure 10 and table 2, we observe that there is a (small) positive correlation between $KnowlTrans^2$ and KnowlTrans; (very small, almost 0) positive correlation between $KnowlTrans^2$ and Risk.

4 Linear Regression with Two Variables

In this section, we will experimenting with linear regression model using a dataset with two features: {KnowlTrans, RespEtiq}. An important fact is that these two features have different data values. *KnowlTrans* values are logit, whereas *RespEtiq* values are based on likert scales (low=1, high =5). So, in order to create a cummulatvive sum plot, we need to normalize the data. Modeling linear regression to predict based on this dataset (KnowlTrans,RespEtiq, Risk), we need to normalize the dataset. (This means, we do not have an option of, whether to perform normalization or not)

We performed Pearson correlation on the normalized dataset, and formulated a contingency table stated in table 3. Seeing the plot in figure 12, it is observable that Risk is negatively correlated with KnowlTrans, Risk is slightly positive correlated with RespEtiq and RespEtiq is positively correlated with KnowlTrans. Even in, linear regression with two features,

Table 3: Contingency table for linear regression with {KnowlTrans, RespEtiq} features in dataset

	KnowlTrans	RespEtiq	Risk
KnowlTrans	1.0	0.2187	-0.0672
RespEtiq	0.2187	1.0	0.0092
Risk	-0.0672	0.0092	1.0

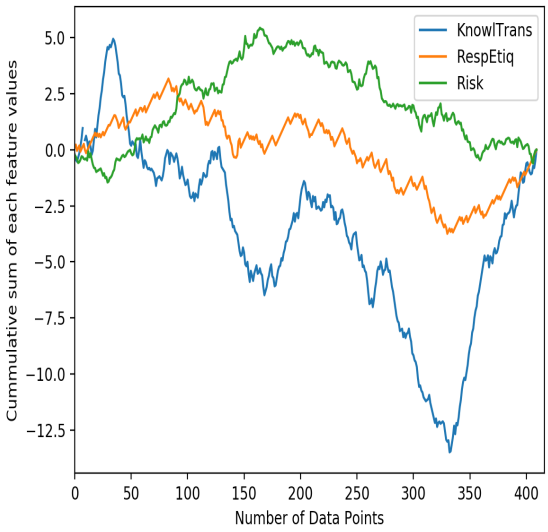


Figure 12: Individual cumulative sum of KnowlTrans, RespEtiq and Risk over all the data points. Data points means observations.

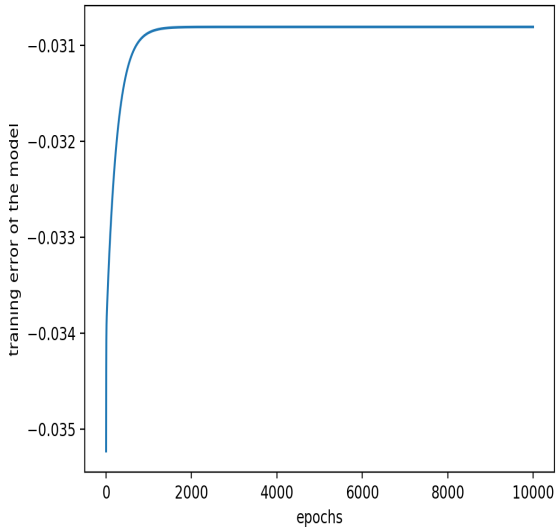


Figure 13: Training error of linear regression model when trained over the dataset with 2 features (KnowlTrans and RespEtiq).

the covariance between features {KnowlTrans, RespEtiq} and *Risk* is -0.02674 (negative correlation).
In the figure 13, there is a smooth asymptotic convergence of the model. This is attributed to small learning rate of the model and less covariance between features. The model was trained over a learning rate of 0.001 for 10000 epochs. We also state that too low learning rate, not necessarily will provide you low prediction error. In the figure 14, we achieve a low prediction error at learning rate of 0.008 of 0.165.

5 Cross Validation Results and Plotting Model Prediction over Observed Risk

We generated the cross-validation results setting the number of epochs to 10000 and learning rate to 0.001. On observing the table 4, we see that linear regression with two variables (KnowlTrans and RespEtiq) performed well over linear regression with one variable and quadratic features. For defining the performance of the model under three different scenarios (LR1V, LR2V, LRQF), we used root mean square error (RMSE) metric. RMSE is one of the most popular performance metric used in the machine learning community to assess the efficacy of regression and recommendation models. For classification models, Receiver Operating characteristics (ROC), Area Under Curve (AUC) and Precision-Recall Curve are the predominant performance metrics. Linear regression model over one feature dataset,

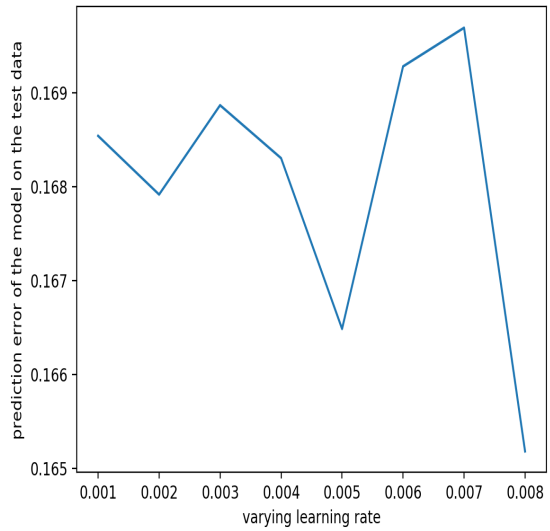


Figure 14: Trend in the prediction error of linear regression model when trained over the dataset with 2 features (KnowlTrans and RespEtiq) with varying learning rate between [0.001, 0.009] with step size of 0.001, epochs were set to 10000

Table 4: Cross Validation results tables. LR1V : Linear Regression with One Variable, LRQF : Linear Regression with Quadratic Features and LR2V : Linear Regression with Two Variables

%age Training	%age Testing	Normalized LR1V	Normalized LRQF	Normalized LR2V
20	80	0.1690	0.1872	0.1753
30	70	0.1670	0.1919	0.1692
40	60	0.1670	0.1727	0.1710
50	50	0.1676	0.1781	0.1670
60	40	0.1660	0.1836	0.1672
70	30	0.1655	0.1711	0.1690
80	20	0.1656	**0.1710**	0.1682
90	10	**0.1650**	0.1721	**0.1649**

provided lowest RMSE of 0.1650, when trained over 90% of the data. The model created for a dataset with features (KnowlTrans and $KnowlTrans^2$) provided lowest RMSE of 0.1710, when trained over 80% of the data. There was a rise of 3% in the prediction error. Finally, when the model was trained and tested over a dataset with features (KnowlTrans, RespEtiq), the lowest error reported was 0.1649, close to first scenario model (LR1V).

In the figure 15, we compare the efficiency of the model when it is trained on a data set with feature *KnowlTrans* and with features *KnowlTrans* and $KnowlTrans^2$. Using the visualization in figure 15 and table 4, it is observable that squaring the feature and adding as training does not lower the error. We observe that fitting of the curve is more smooth in figure 16a rather than figure 16b. Furthermore, we enhance our understanding by comparing the prediction of model over all the three scenarios.

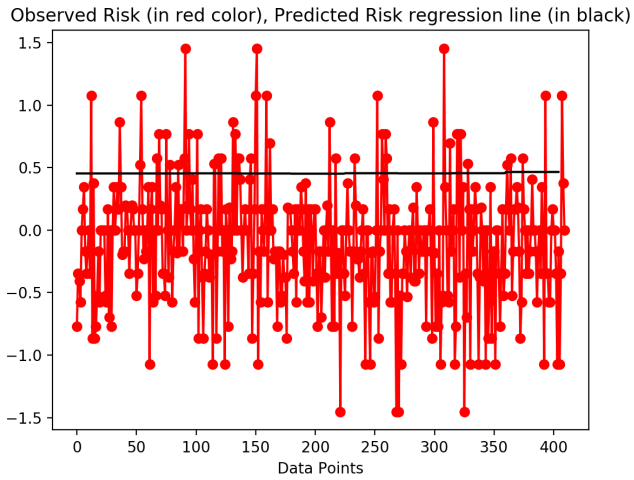
Using the table 4 and figure 16, we identified following points :

- The linear regression model performed well when it used two features (KnowlTrans, RespEtiq) and training data was 90% and testing data was 10%.
- There was very small difference between the coefficient values of $\theta's$ for KnowlTrans and RespEtiq. (KnowlTrans = -0.002 and RespEtiq = 0.03).
- There was significant difference between the coefficient values of $\theta's$ for KnowlTrans and $KnowlTrans^2$. (KnowlTrans = -0.1125 and $KnowlTrans^2 = 0.1622$)
- Pearson chi-square statistic applied over the dataset (KnowlTrans and $KnowlTrans^2$) provided a p-value < 0.05 showing that both the features are important for predicting risk.
- Pearson chi-square statistic applied over the dataset (KnowlTrans and RespEtiq) provided a p-value > 0.05 (nearly 1) showing that RespEtiq does not any impact on the prediction of risk. This can be shown in the table 4.
- There is small amount of overfitting seen in the regression line in figure 16b, where as there is no overfitting in figure 16a and 16c.
- The predicted risk lies around 0.5 in figure 16a and 16b whereas, the predicted risk is centered around 0.0 in figure 16c. This behavior in figure 16c is attributed to RespEtiq, which is a likert scale based feature, weighting each observation in KnowlTrans.

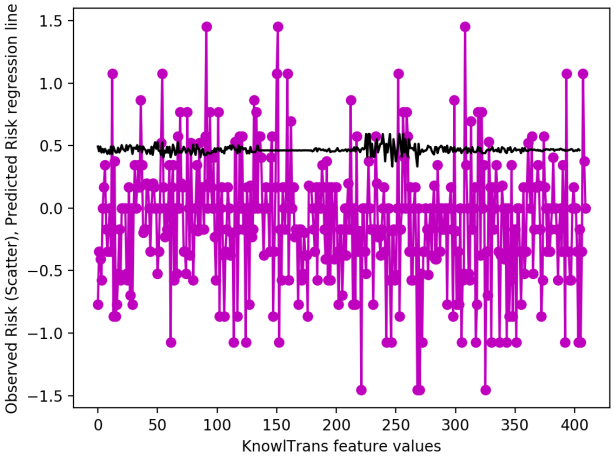
Reference

[1] Aggarwal, Charu C., and ChengXiang Zhai, eds. Mining text data. Springer Science and Business Media, 2012.

Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

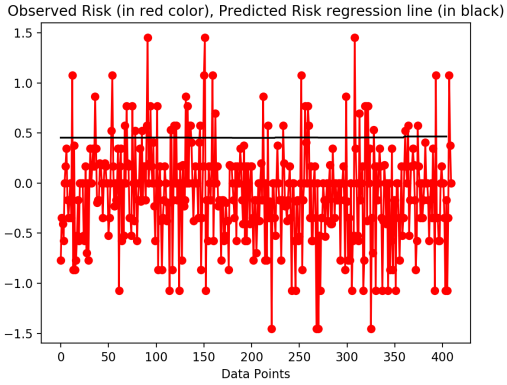


(a) LR1V

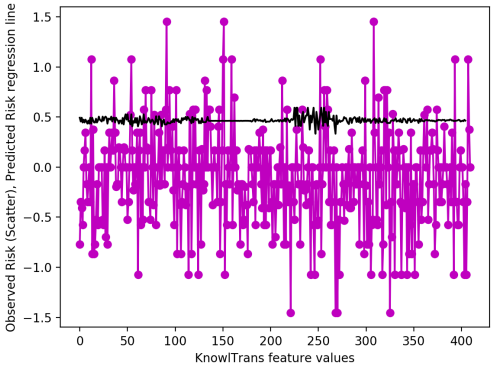


(b) LRQF

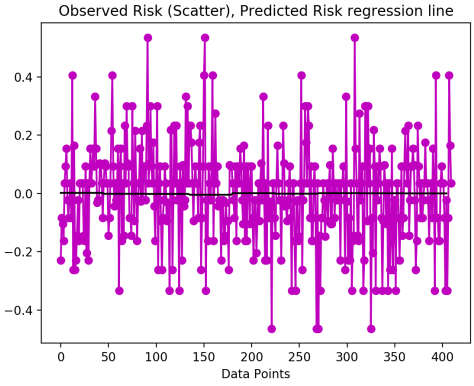
Figure 15: (a). Observed Risk and Predicted Risk plot when the linear regression model was trained over one feature (KnowlTrans). Data points means observations. (b). Observed Risk and Predicted Risk plot when the linear regression model was trained over KnowlTrans and $KnowlTrans^2$. `KnowlTrans feature values` means observations.



(a) LR1V



(b) LRQF



(c) LR2V

Figure 16: (a). Observed Risk and Predicted Risk plot when the linear regression model was trained over one feature (KnowlTrans). (b). Observed Risk and Predicted Risk plot when the linear regression model was trained over KnowlTrans and $KnowlTrans^2$. `KnowlTrans feature values` means observations. (c). Observed Risk and Predicted Risk plot when the linear regression model was trained over two features viz. KnowlTrans and RespEtiq. Data points means observations.