TRAIN DELAY PREDICTOR

CSL341: COURSE PROJECT

ABDUL HADI SHAKIR
ANUBHAV SINGH
BHARAT RATAN
RAHUL NAGAR

Indian Railways

- Lifeline of India
- **25 million** passengers daily
- **10,000** trains running daily
- Running schedule often gets *derailed*
- Lets help out the passengers…

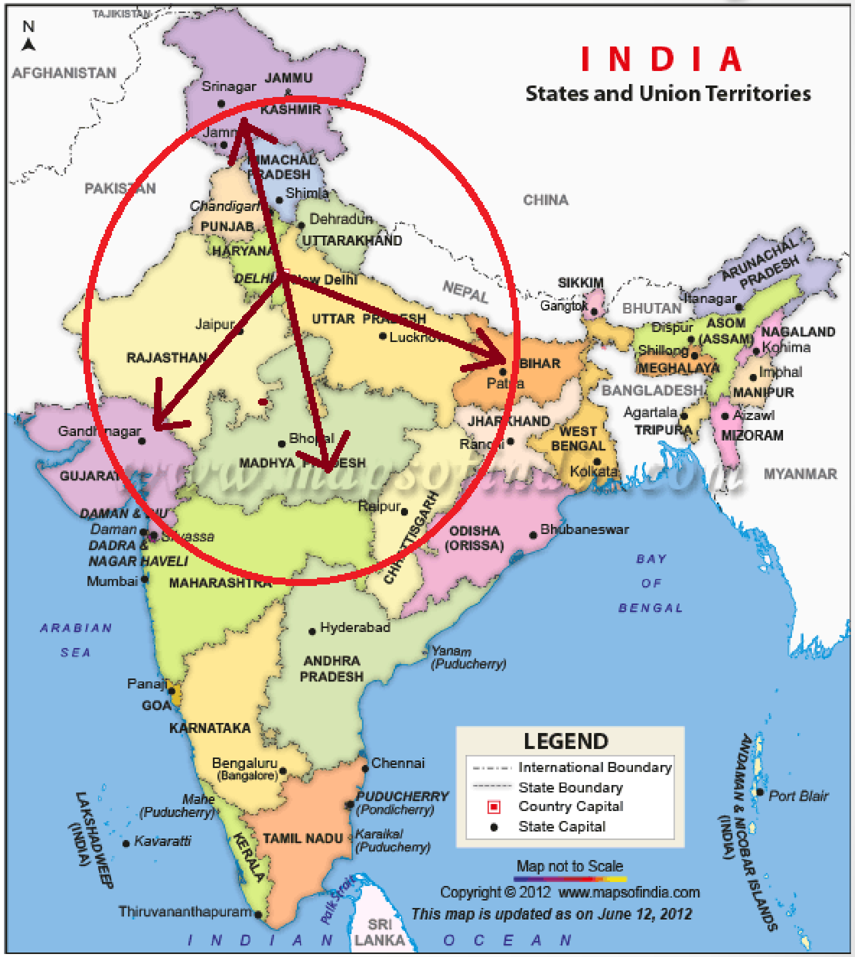
Problem Definition

Deploy Machine Learning to predict the delay in arrival of train(s), using features:

- Type of Train
- Travelling Distance
- Day of Week
- Region of Journey

Modeling Indian Railway

- Identify major station in IRN in terms of degree of station (connecting stations) and weight of station (traffic).
- Delhi/NCR as the origin.
- Destination (zones):
  - Southern UP/Madhya Pradesh Region
  - Eastern UP/Bihar Region
  - Rajasthan/Gujarat Region
  - Chandigarh/Jammu Region



Data Collection

- Automated data collection from *runningstatus.in*
- Used python cgi-based scripts
- Collected for over a month
- 80% data points for training
- 20% data points for testing

Models we looked upon

- **Support Vector Machine (SVM)**
- Polynomial Regression
- Random Forest

SVM

- Used ***libsvm*** tool for matlab
- Gaussain/RBF kernel with different parameters
- Experimented with various SVM models
- *Binary clasification viz.*
  - *Delayed (> 15 mins).*
  - *Non-Delayed (0-15 mins)*

SVM..

- Overfitting of data
- Good on train, bad on test
- Accuracies
    - Train : 87.16%
    - Test : 72.12%
- Why?
- Variety and variance of parameters
- Narrowing the parameters didn’t help

Models we looked upon..

- Support Vector Machine (SVM)
- **Polynomial Regression**
- Random Forest

Polynomial Regression

- Used Linear Regression with Polynomial Basis Functions
- varied *d* from 1 to10 and calculated training and test errors.
- Training error did not converge for $d > 3$
- Discarded the model

Models we looked upon..

- Support Vector Machine (SVM)
- Polynomial Regression
- **Random Forest**

Random Forest

- An *ensemble learning* method
- Grows multiple *decision trees* (forms *forest*)
- Uses *voting* for classification

Random Forest..

- Grouped data into *bins* (for classification)
  - 0-3 min = bin-1
  - 4-10 min = bin-2
  - 11-25 min = bin-3
  - 25-60 min = bin-4
  - > 60 min = bin-5
- Each decision tree trained on ⅔ of the data *(bag)*
- Tested on remaining ⅓ of data *(out of bag)*
- Error metric
  - OOB_{error} = total error averaged over forest
- Satisfying results
  - Training Accuracy = 84.21%
  - Test Accuracy = 82.37%

Model Selection

MODEL	TRAINING ACCURACY	TEST ACCURACY
SVM	87.16%	72.12%
RANDOM FOREST	84.21%	82.37%

SVM – Overfits training data ☒

Random Forest – Equally well in training and test ☑

Analysis

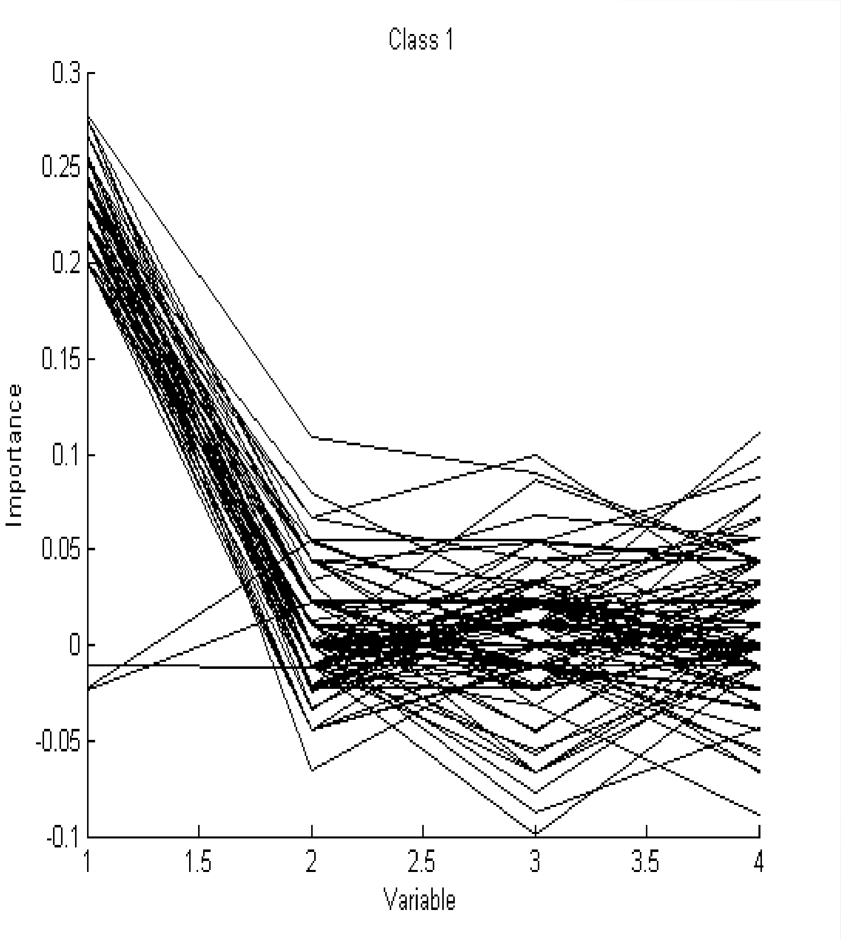
- **Variable Importance**
- Variable Interaction

Variable Importance

- V1 = #votes for correct class using OOB cases
- Randomly permute the values of variable '*m*' in the OOB cases
- V2(m) = #votes for correct class using permuted OOB cases
- Importance(m) = (V1 – V2(m)) averaged over forest
- High values suggest that the corresponding variable is important in correctly classifying the case

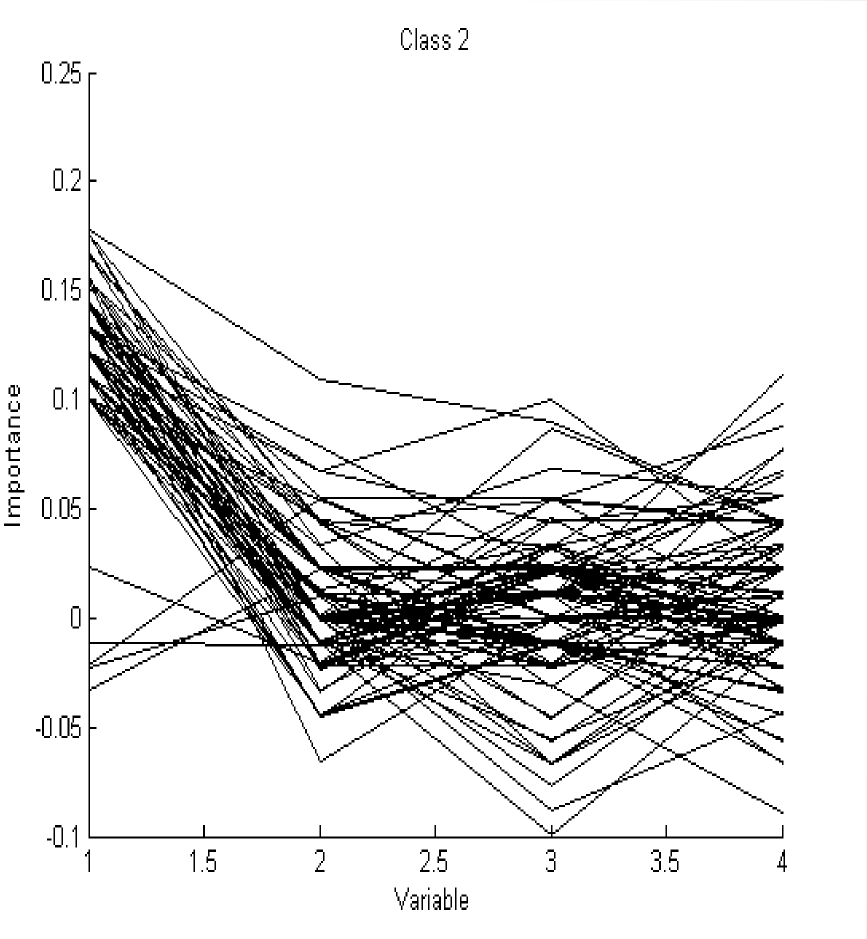
Variable Importance..

- For Class-1
- Variable-1 is important
- Class-1 = 0-3 min delay
- Variable-1 = train type
- Conclusion:
  - Certain trains never get delayed
  - High priority trains
  - Rajdhani, Shatabdi etc.



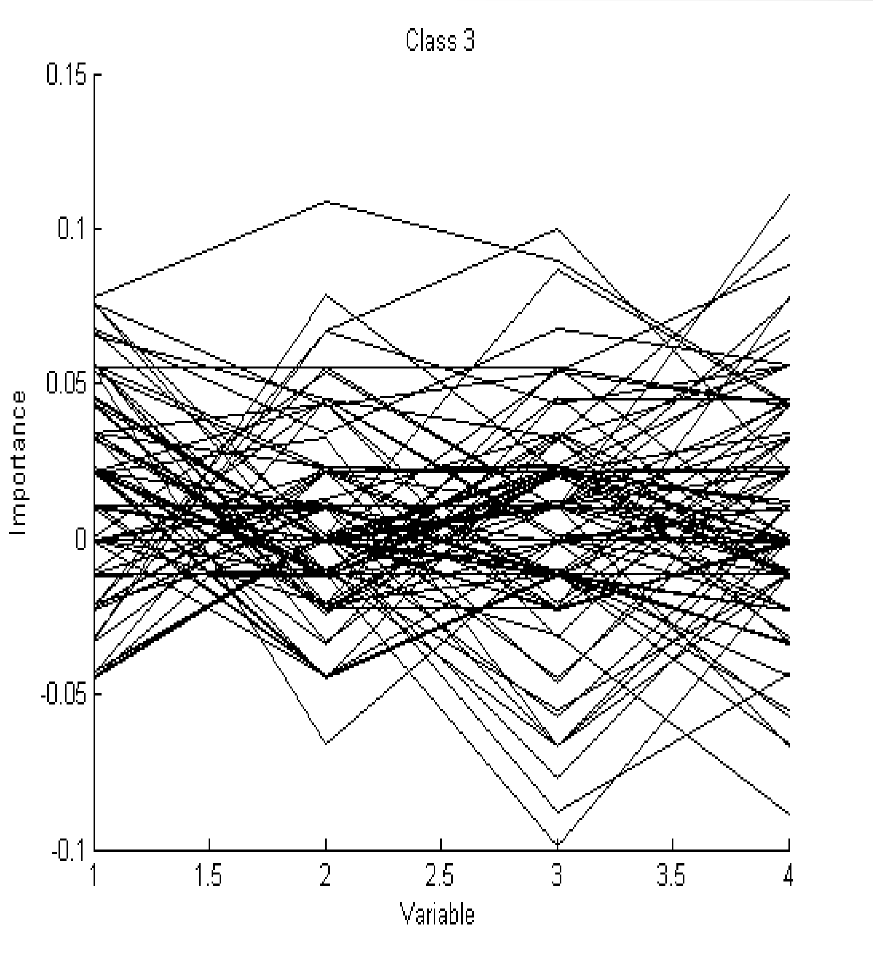
Variable Importance..

- For Class-2
- Variable-1 is important
- Class-2 = 4-10 min delay
- Variable-1 = train type
- Conclusion:
  - Certain trains only slightly get delayed
  - High priority trains
  - Rajdhani, Shatabdi etc.



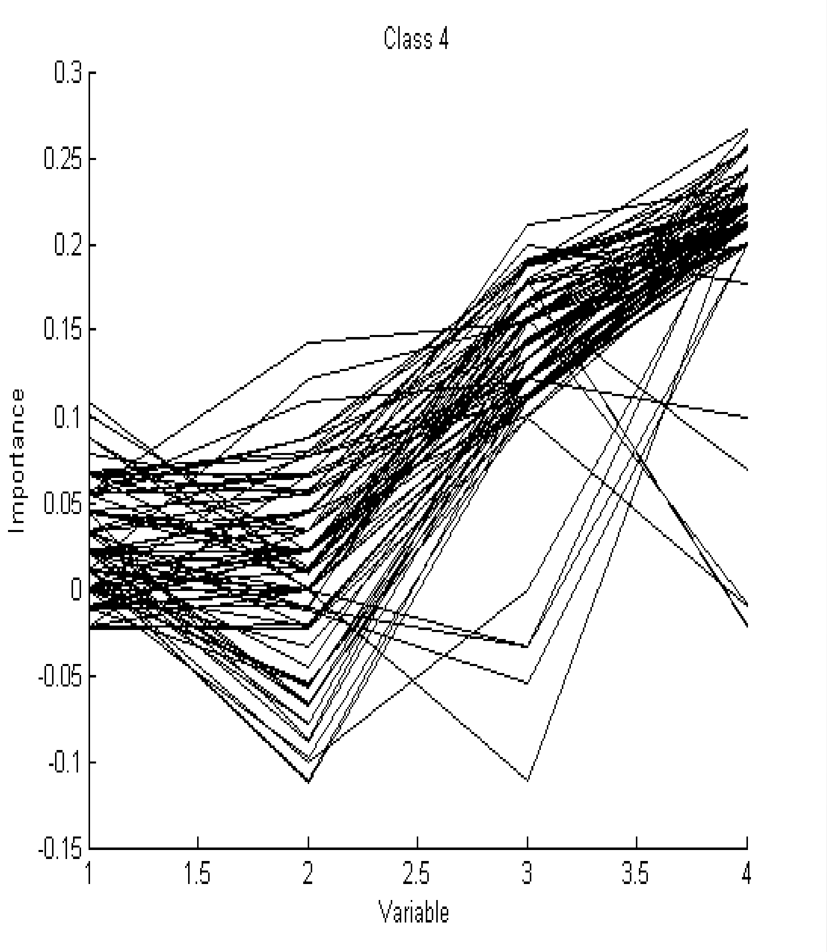
Variable Importance..

- For Class-3
- All variables have similar importance
- Class-3 = 11-25 min delay
- Nothing much can be concluded



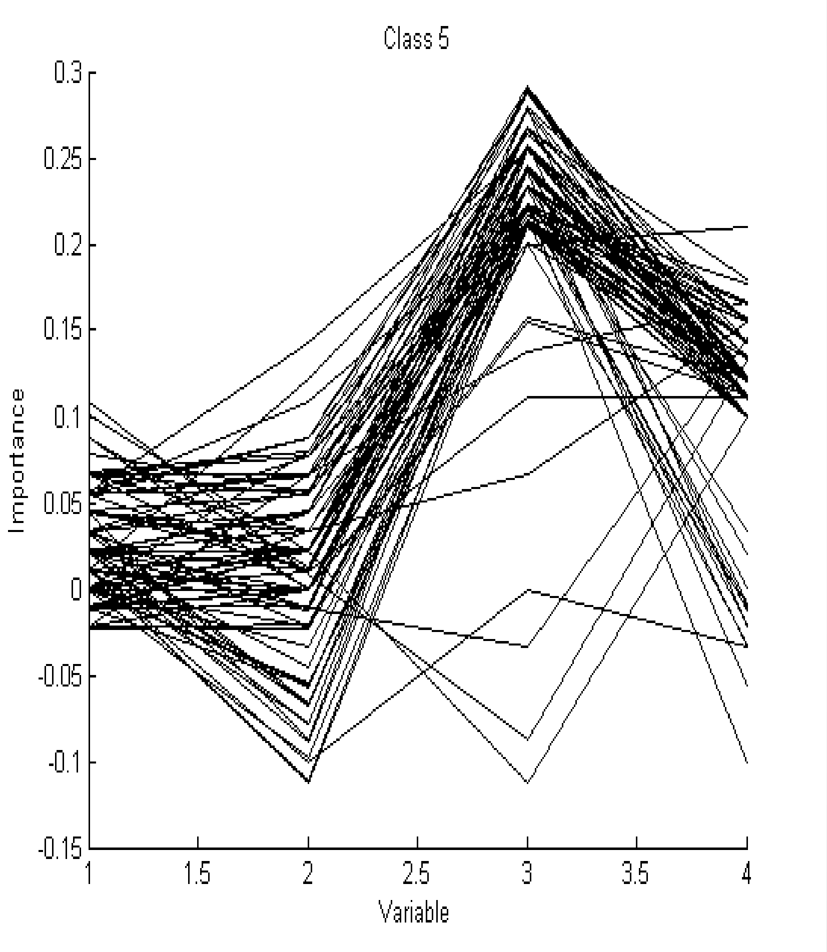
Variable Importance..

- For Class-4
- Variable-4 is important
- Class-4 = 25-60 min delay
- Variable-4 = Day of week
- Conclusion:
  - Trains are delayed on specific days of week
  - Weekends
  - Region of run also important



Variable Importance..

- For Class-5
- Variable-3 is important
- Class-5 = > 60 min delay
- Variable-3 = Region of run
- Conclusion:
  - Trains running in certain regions are drastically late.
  - UP/Bihar zone



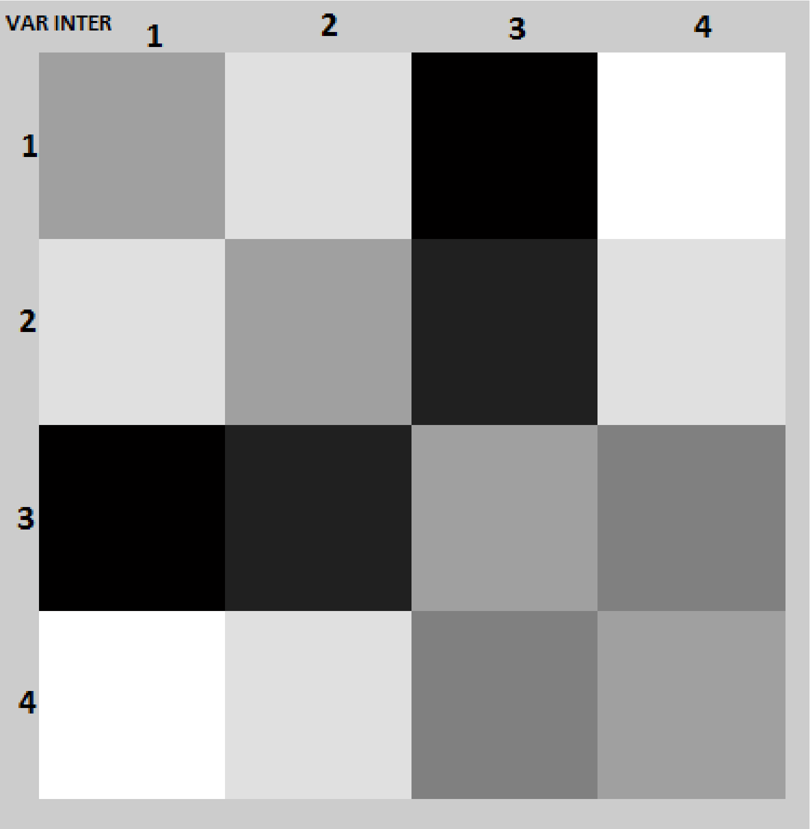
Analysis..

- Variable Importance
- **Variable Interaction**

Variable Interaction

- Variables ‘*m*’ and ‘*k*’ interact if a split on one variable, say ‘*m*’, in a tree makes a split on ‘*k*’ either systematically less possible or more possible
- Absolute difference of gini values averaged over the forest

Variable Interaction..



Strong interaction between (**var-1,var-3**) and (**var-2,var-3**)

Variable Interaction..

- (var-1,var-3) = (train type, region of run)
- (var-2,var-3) = (distance, region of run)
- Conclusion:
  - Trains of certain priority – no matter in which region they run it never get delayed (like Rajdhani).
  - Trains in certain regions always get delayed (UP/Bihar) irrespective of whether they are long or short distance trains.

Future Scope

- X-Factor
- Widening the scope of Regions/Stations
- Using Railway Infrastructure
- Let’s look at it from other side

QUESTIONS ?

THANK YOU