GetMyFlight

Cheapest Flight Rates in No Time

Group No - 3
Aastha Grover
Ankur Bag
Neelesh Saxena
Tushar K

Problem Statement

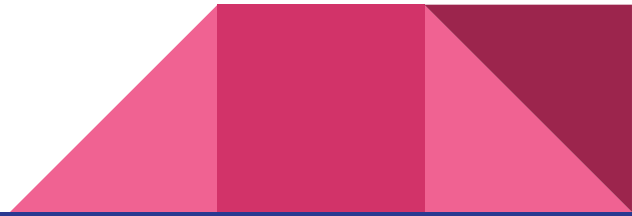
- Unpredictability and uncertainty of flight fares.
- ‘The earlier you book, the cheaper you get’ - is always not true.
- Flight Reservation websites are in a rush to provide their customers with the cheapest flights.
- Nobody is bothered how the rates will vary in the future. This is where our research is centered.

Our Proposal/Suggestion:

- Build a flight recommendation system -
  - To predict the flight rates to a particular destination for the given dates.

Synopsis

- Actor/Use Case
- Architecture & Infrastructure
- Data Preprocessing Phase
- Algorithm Selection Phase



Actor/Use Cases

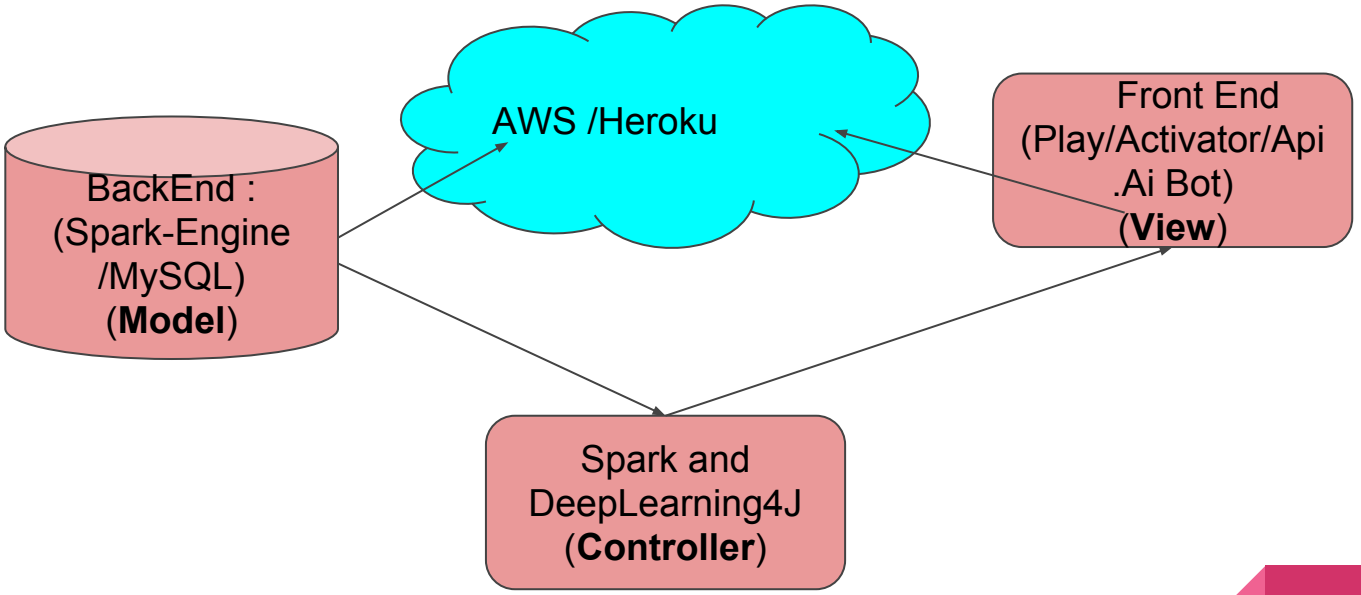
Actor

User is the sole actor of the system.

Use Case

- ➔ User will input Preferred dates and Source-Destination.
- ➔ System will predict when to book the tickets.

Architecture



Infrastructure

- **User Interface**: PlayFramework , Scala Controllers, Javascript, Bootstrap,CSS, HTML and api io.
- **DataBase** : In memory database (PostgresSQL), MySql deployed on Amazon Web Services, Slick for Database Querying.
- **Apache Spark - Scala Integration :** Neural Network Algorithm implemented in Scala using deeplearning4j.
- **Data Cleansing & Preparation** : R Script

Data Preprocessing Phase

About Data

- Data is for two years (2015 and 2016) and it was originally divided into quarters with each quarter having approximately 50000 records and 26 columns.
- Data has several features like : CARRIER, ORIGIN, DESTINATION, SEATS (available), DATE_OF_BOOKING, DATE_OF_TRAVEL, BASE_FARE, TICKET_FARE and (LATITUDE,LONGITUDE) for Origin and Destination.
- We gathered the data from different tables which was available on (Official Beureau Of Transportation)

Source : http://www.transtats.bts.gov/

Feature Vector

- **Source, Destination**
- **Latitude & longitude** for source and destination.
- **Number of seats** available for each carrier, as a measure of demand.
- **Date of booking** for flight.
- **Base Fare** : It is a fixed value set by International Air Transport Association(IATA).
- **Date Of Travel**
- **Ticket Fare:** Market fare for each itinerary.

Feature Selection

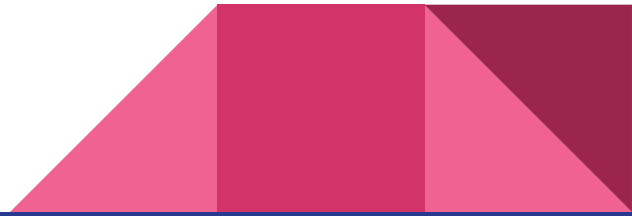
- We used linear regression and P-values statistic to identify the correlation of each input feature with the predicted variable (Ticket Fare).
- We also used wrapper methods to try different combinations of attributes.
- Tested the attribute combinations with forward and backward passes to add or remove features which gave us the best combination of input attributes.
- The final attributes used for building a prediction model are:

I. Carrier
II. Base Fare
III. Origin-Destination
IV. Date of Travel
V. Seats Available

Algorithm Selection Phase

Possible Machine learning Algorithms

- Multiple Linear Regression
- Decision Trees
- Neural Networks



Data Preparation For Neural Network (R Script)

- Neural Networks gives optimized performance only on the normalized data inputs. Range between -1 to 1 is desired although higher numeric ranges also work depending on the feature set.
- Therefore we normalized the categorical & continuous variables using different techniques suitable for each column.

1. **Categorical Variables** : Assigned their occurrence frequency and used Decimal - Binary Conversion.
2. **Continuous Variables** : Z-Score normalization.

Implementing Neural Network

1. **Set the Parameters** (Seed, Train Sample, learning Rate, Number of hidden Layers, Number of nodes in each hidden layer).
2. **Feature Selection** to predict the Ticket fare.
3. **Training** : Build the network of neurons Using Multilayer Network class of deeplearning4j.
4. **Tuning Parameters**: Adjusted hidden layers, number of nodes per hidden layer and learning rate to improve the prediction.
5. **Testing** : Predicting the Ticket fare (Target Variable) for 2016 dataset.

Conclusion:

- Lower the learning rate , higher the prediction accuracy.
- Adjusted the parameters to avoid overfitting.

Evaluation

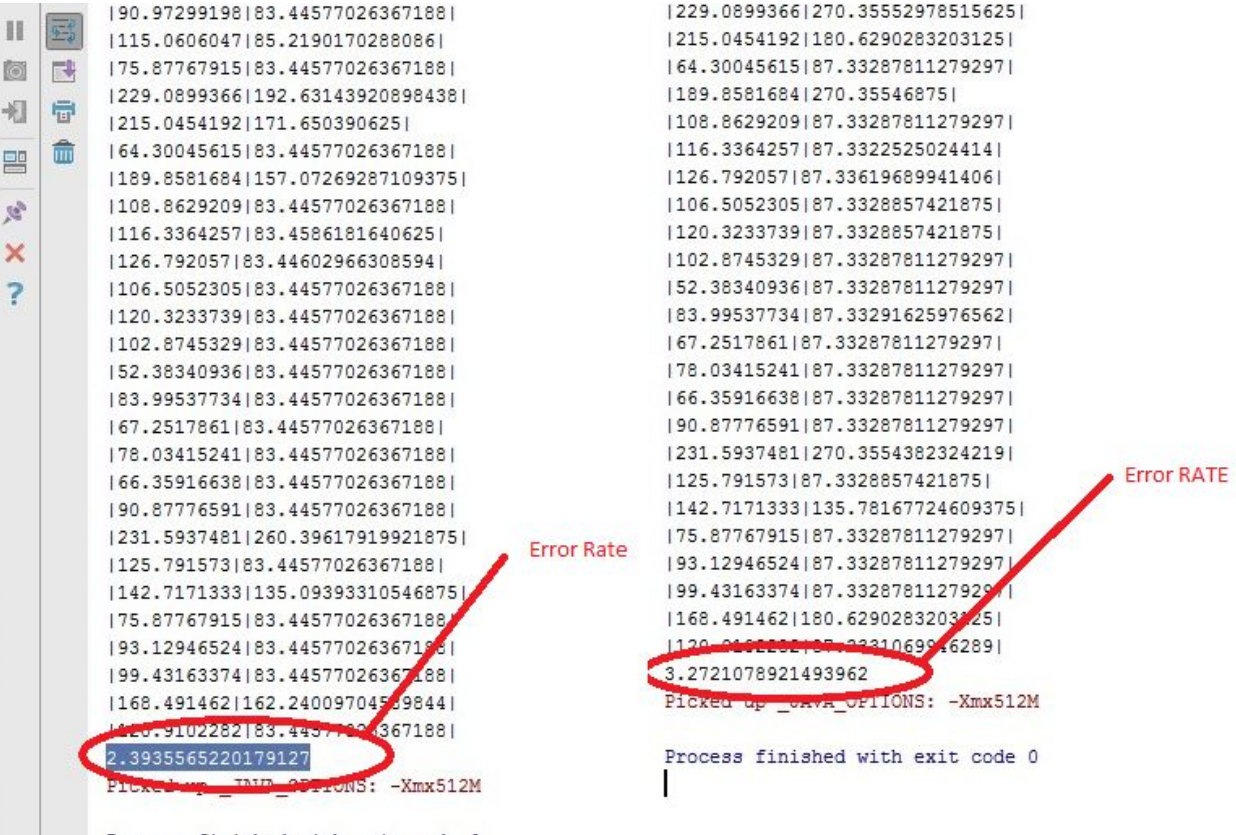
Measuring Prediction Accuracy

RMSE

Root-mean-square error (**RMSE**) is used to measure the differences between values predicted by a model o and the values actually observed.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Error Rate Across Three Runs



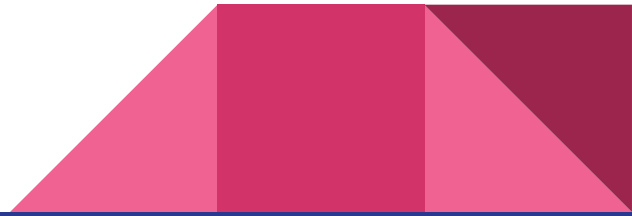
Application and User Interface

FLIGHT DATABASE

Search Source	Search Destination	Search Month of Travel	Search Day of Travel	Carrier	Actual Price	Predicte Price
Aguadilla PR	Newark NJ	1	1	United Airlines	181.98493	263.4828
Aguadilla PR	Newark NJ	1	1	United Airlines	216.4885	263.5129
Akron OH	Atlanta GA	1	1	Southwest Airlines	66.3355	92.3116
Akron OH	Atlanta GA	1	1	Southwest Airlines	87.90023	92.3116
Akron OH	Orlando FL	1	1	Southwest Airlines	124.27969	92.3116
Akron OH	Washington DC	1	1	Southwest Airlines	93.12947	92.3116
Akron OH	Washington DC	1	1	Southwest Airlines	88.81652	92.3116

Using Play Framework

- Object Modeling
- Database Selection
- Created Test specs for Application, Scala Controllers and repository.
- Json, Javascript, DataTables



Using api.ai

- Conversational User Interface
- Bot Application to ask System the prediction

http://getmyflight.mx7a8jdfvi.us-west-2.elasticbeanstalk.com/

Acceptance Criteria

Proposed Criteria	Actual Criteria
Number Of Carrier (4)	Number Of Carrier (13)
Number Of Cities to be Analysed(6)	Number Of Cities to be Analysed(more than 20)
3 successful predictions out of 5.	Accuracy of Prediction Model :70%

Challenges Faced & Solutions

1) Data Collection issues : Merged the data from different tables
2) Unpredictable results with initial feature vector.
3) Technology to use since the data was too big : 365 days
4) Best Suitable Algorithm for the given problem at hand.
5) Restructuring the data to use it for input format.
6) Customizing the algorithm
7) Normalizing the data (Categorical a well as Continuous)for input to neural network.
8) Application configuration issues with Play Framework, Data Modelling

Scope

- Integrating Play Framework application with our Spark algorithm so that we can make near real time predictions.
- Predicting for International as well as Domestic Flights.
- Taking Roundtrips into consideration.

Milestones

Key Milestones	Start Date	End Date
A. Feature Selection B. Dataset Creation. C. Decide the architecture of the application.	11/3/2016	11/10/2016
A. Data Analysis - Data Visualization, Data Cleansing / Manipulation B. Decision on the machine learning algorithm.	11/10/2016	11/17/2016
A. Build the Predictive algorithm/Model.	11/17/2016	11/24/2016
A. Application Integration , Backend Complete	11/24/2016	12/5/2016
A. Application Test Run Executed, FrontEnd Completed	12/1/2016	12/9/2016

Code Repository

https://github.com/ankurbag/CSYE7200_Scala_Project_Group3

Thank You :)