

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

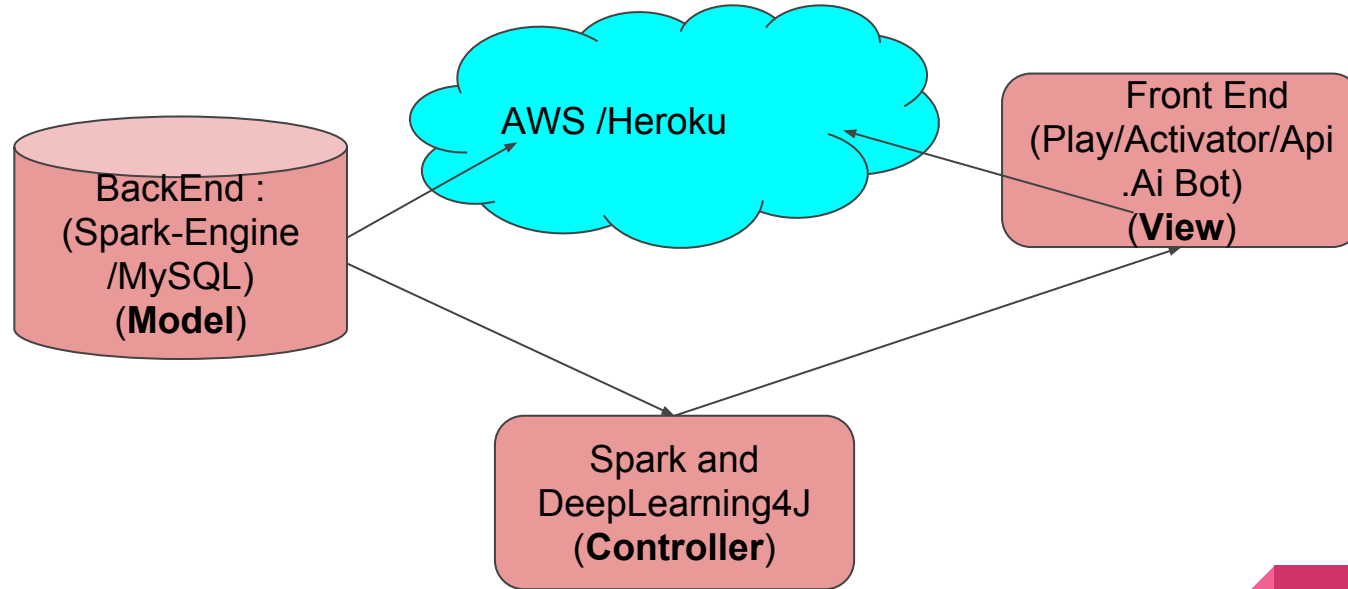
User is the sole actor of the system.

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 (PostgreSQL), 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 [\(Official Beureau Of Transportation\)](#)

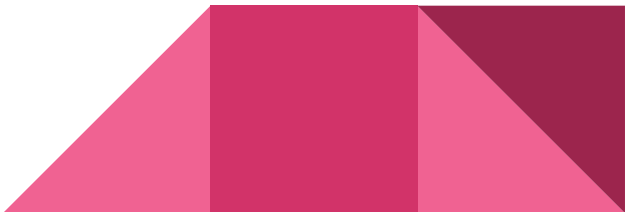
Source : <http://www.transtats.bts.gov/>



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 \hat{y}_i and the values actually observed y_i .

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



Error Rate Across Three Runs

The image displays three screenshots of Java debug output, each showing a list of error rates. A red circle highlights a specific error rate in each screenshot, and a red arrow points to it with the label "Error Rate".

Run 1 (Left):

```
|90.97299198|83.44577026367188|
|115.0606047|85.2190170288086|
|75.87767915|83.44577026367188|
|229.0899366|192.6314392089438|
|215.0454192|171.650390625|
|64.30045615|83.44577026367188|
|189.8581684|157.07269287109375|
|108.8629209|83.44577026367188|
|116.3364257|83.4586181640625|
|126.792057|83.44602966308594|
|106.5052305|83.44577026367188|
|120.3233739|83.44577026367188|
|102.8745329|83.44577026367188|
|52.38340936|83.44577026367188|
|83.99537734|83.44577026367188|
|67.2517861|83.44577026367188|
|78.03415241|83.44577026367188|
|66.35916638|83.44577026367188|
|90.87776591|83.44577026367188|
|231.5937481|260.39617919921875|
|125.791573|83.44577026367188|
|142.7171333|135.09393310546875|
|75.87767915|83.44577026367188|
|93.12946524|83.44577026367188|
|99.43163374|83.44577026367188|
|168.491462|162.24009704589844|
|120.9102282|83.44577026367188|
2.3935565220179127
Picked up _JAVA_OPTIONS: -Xmx512M
Process finished with exit code 0
```

Run 2 (Middle):

```
|229.0899366|270.35552978515625|
|215.0454192|180.6290283203125|
|64.30045615|87.33287811279297|
|189.8581684|270.35546875|
|108.8629209|87.33287811279297|
|116.3364257|87.3322525024414|
|126.792057|87.33619689941406|
|106.5052305|87.3328857421875|
|120.3233739|87.3328857421875|
|102.8745329|87.33287811279297|
|52.38340936|87.33287811279297|
|83.99537734|87.33291625976562|
|67.2517861|87.33287811279297|
|78.03415241|87.33287811279297|
|66.35916638|87.33287811279297|
|90.87776591|87.33287811279297|
|231.5937481|270.3554382324219|
|125.791573|87.3328857421875|
|142.7171333|135.78167724609375|
|75.87767915|87.33287811279297|
|93.12946524|87.33287811279297|
|99.43163374|87.33287811279297|
|168.491462|180.6290283203125|
|120.9102282|87.33291625976562|
3.2721078921493962
Picked up _JAVA_OPTIONS: -Xmx512M
Process finished with exit code 0
```

Run 3 (Right):

```
|189.8581684|267.7142028808594|
|108.8629209|89.85895538330078|
|116.3364257|89.85811614990234|
|126.792057|89.85895538330078|
|106.5052305|89.85895538330078|
|120.3233739|89.85895538330078|
|102.8745329|89.85895538330078|
|52.38340936|89.85895538330078|
|83.99537734|89.85895538330078|
|67.2517861|89.85895538330078|
|78.03415241|89.85895538330078|
|66.35916638|89.85895538330078|
|90.87776591|89.85895538330078|
|231.5937481|203.98495483398438|
|125.791573|89.85895538330078|
|142.7171333|89.9111099243164|
|75.87767915|89.85895538330078|
|93.12946524|89.85895538330078|
|99.43163374|89.85895538330078|
|168.491462|203.98495483398438|
|120.9102282|89.85895538330078|
3.301117104614892
Picked up _JAVA_OPTIONS: -Xmx512M
Process finished with exit code 0
```

Application and User Interface



FLIGHT DATABASE

Search Source	Search Destination	Search Month of Travel	Search Day of Travel	Carrier	Actual Price	Predicted 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

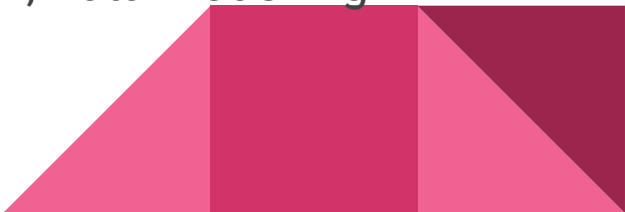
<http://getmyflight.mx7a8jdfvi.us-west-2.elasticbeanstalk.com/>



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 as 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 :)