Introduction

Diana Shen

A Little about Myself

THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

DMA
DALLAS MUSEUM OF ART

DataScience@SMU

On-Time Performance Forecast Project

Diana Shen

What Is On-Time Performance?

- On-time performance (OTP) is one the of key KPIs that are closely monitored in the airline industry. It measures the ability of an airline to be operated on time. Specifically, it is a metric calculating the percentage of on-time flights remaining on the published schedule.
- The OTP used in this project is a daily metric, determined by the number of arrival on-time flights vs. total number of scheduled flights per day, following the 15-minute on-time rule.
  - OTP = (total on-time flights/total scheduled flights)* 100%
- 15-minute on-time rule: a flight that arrives within 15 minutes of its scheduled arrival time is considered on time

Project Background

- My team: a data science/operation research consultancy group
- My internal customers/stakeholder
- Analytics in airline operations: challenges and opportunities
- Opportunities and threats for our team
- Lineup for this project

Why Predicting OTP?

- POC: breaking the ground for predictive analytics in airline operation using weather
- Strategic reasons
- Providing decision-driven insights

DataScience@SMU

Data Cleaning and Wrangling

Diana Shen

Data Overview

- What factors could impact on-time performance?


Seasonality


Weather


Passenger


Crew

- Data considered in this project:
  - Weather forecast data
  - Passenger data: load factors
  - Schedule data: daily number of scheduled flights
  - Seasonality: day of week, month
  - Yesterday’s OTP

Working with Weather Data

- Multiple weather forecast data sources to choose from:

NOAA

WEATHER UNDERGROUND

AccuWeather

- Things to consider when choosing the most suitable weather forecast source:
  - Data quality
  - Data availability
  - Data discrepancy
  - Easiness to obtain

Working with Weather Data

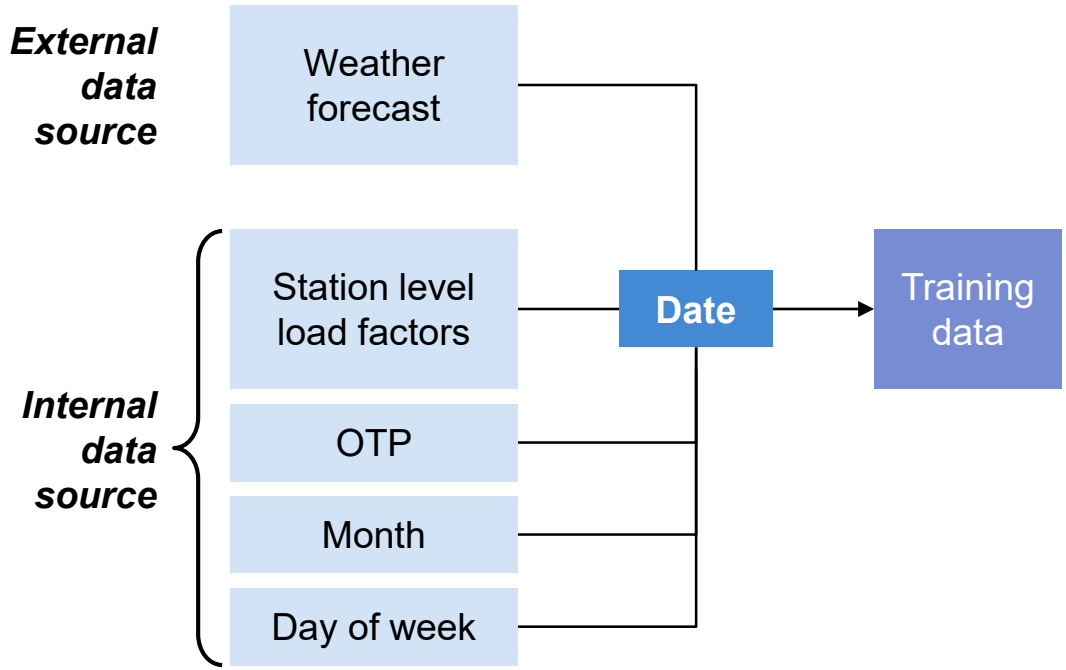
- An example of weather prediction data for one station
- Efforts in understanding the numerous elements in weather forecast data
  - Research
  - Ask domain experts
  - Exploratory analysis

Sample Message

```
KDEN   GFS MOS GUIDANCE    3/04/2010  1200 UTC
DT /MAR   4/MAR   5                /MAR   6                /MAR   7
HR   18 21 00 03 06 09 12 15 18 21 00 03 06 09 12 15 18 21 00 06 12
N/X                    27          53          28          51    24
TMP  48 52 51 40 35 32 30 35 47 50 48 40 38 35 32 36 45 48 45 31 27
DPT  25 25 27 29 29 27 25 25 21 17 18 20 21 21 20 21 21 22 22 23 21
CLD  BK SC SC SC SC OV OV OV BK BK BK SC BK SC SC SC SC SC SC SC SC
WDR  19 14 13 16 20 20 25 25 28 28 28 22 23 23 22 21 12 10 10 17 21
WSP  05 12 11 08 08 07 07 07 08 10 11 08 08 07 07 07 07 10 10 08 07
P06         1     6    14    20    14    25     8     4     2  3  2
P12                    15          27          26           4     3
Q06         0     0     0     0     0     0     0     0     0  0  0
Q12                     0           0           0           0     0
T06      5/11  1/ 2  0/ 1  1/ 2  8/ 3  1/ 1  0/ 0  0/ 0  2/ 8  0/ 0
T12            7/11        1/ 2        8/ 4        0/ 1     2/ 8
POZ   1  1  3  2  2  7  8  3  4  2  3  0  3  3  4  6  3  1  2  0  3
POS  27 16 18 24 28 65 74 72 31 20 37 57 94 87 95 65 41 18 28 50 91
TYP   R  R  R  R  R  S  S  S  R  R  R  S  S  S  S  S  R  R  R  R  S
SNW                     0                          0                0
CIG   8  8  8  8  8  8  8  7  8  8  7  8  8  8  8  8  8  8  8  8  8
VIS   7  7  7  7  7  7  7  1  7  7  7  7  7  7  7  7  7  7  7  7  7
OBV   N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N
```

Data Blending

- Querying external data: Python web scraping
- Querying Internal data: SQL
- Blending: Python
- Remember to do sanity check



DataScience@SMU

Modeling

Diana Shen

Overview of OTP Model

Seasonality

- Considering the operation differences among **days of week** and **months**

Weather

- Weather information includes visibility, wind speed, snow duration and etc. **per day**
- Using **historical** weather of **20** stations at **different time of day** (morning, afternoon, evening, and midnight)

Load factor

- Using **historical station-level** average load factor **per day**
- It includes load factors of **20 stations** each day

OTP

- Acting as a **target** for the model to learn from
- Using **daily** OTP data of **three years**

Predictive modeling

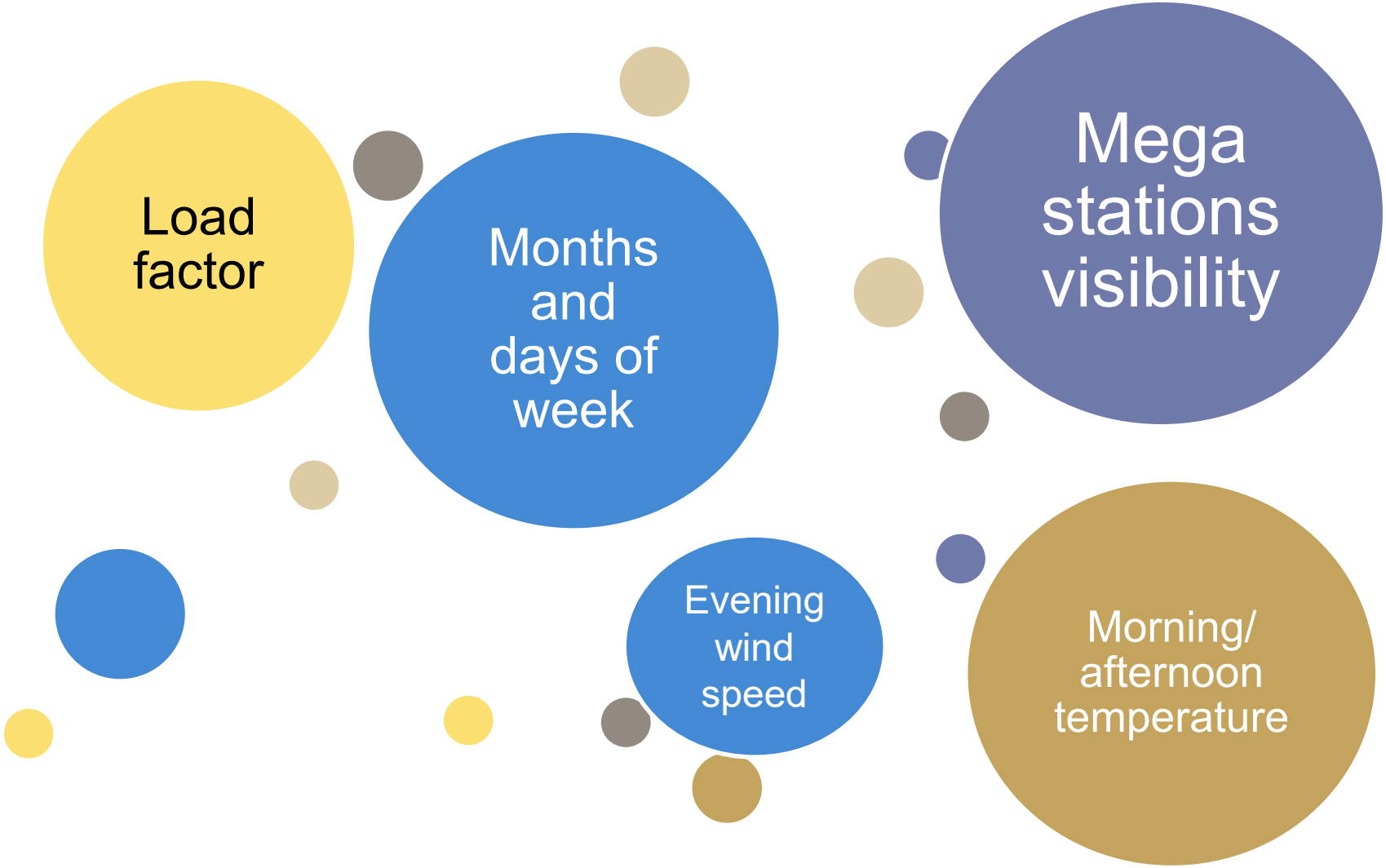
Machine learning

- Find patterns within data
- Extract data features that well correlate/represent OTP using a model

Machine Learning Models Investigated

Tree-based models	Gradient Boosting Model
	Random Forest
Linear models	Multivariate Linear Regression
	Lasso
	Ridge
Artificial intelligence	Neural Network

Influential Variables in the Model



* Bubble size is proportional to influential level

Cross Validation

- Variables used vs. data points available for training
  - Why three years of training data?
- Benefits of using cross validation
- **RMSE**

Final Model

- Factors considered when choosing a model:
  - Performance
    - Model accuracy
    - Dimension reduction
  - Interpretability
  - Time needed to run
- Lasso is the final model implemented in this project

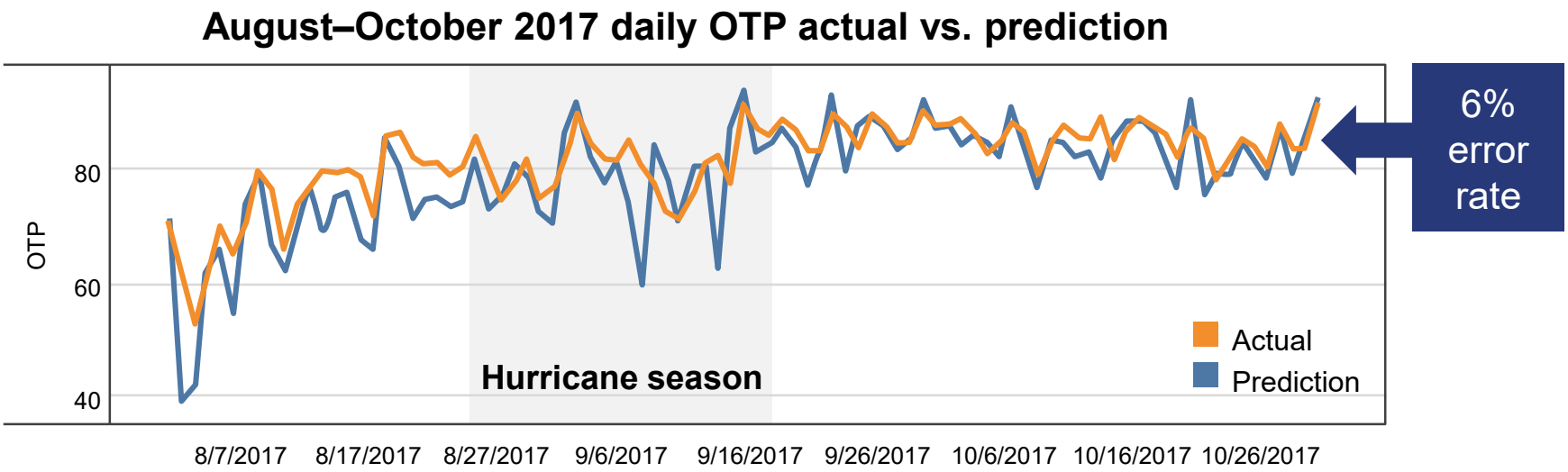
DataScience@SMU

Visualization

Diana Shen

Visualization on Model Performance

What plot could be used? Q-Q plot? Scatterplot?



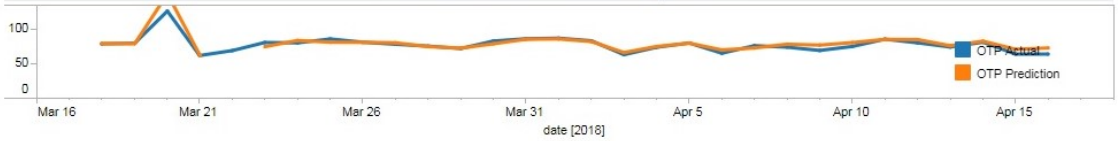
Dashboard Presented to Clients

OTP Prediction for Tuesday, April 17, 2018

Lower Bound	OTP Prediction	Upper Bound
77.877	82.601	87.006

A prediction with confidence intervals

OTP Actual/Prediction Comparison



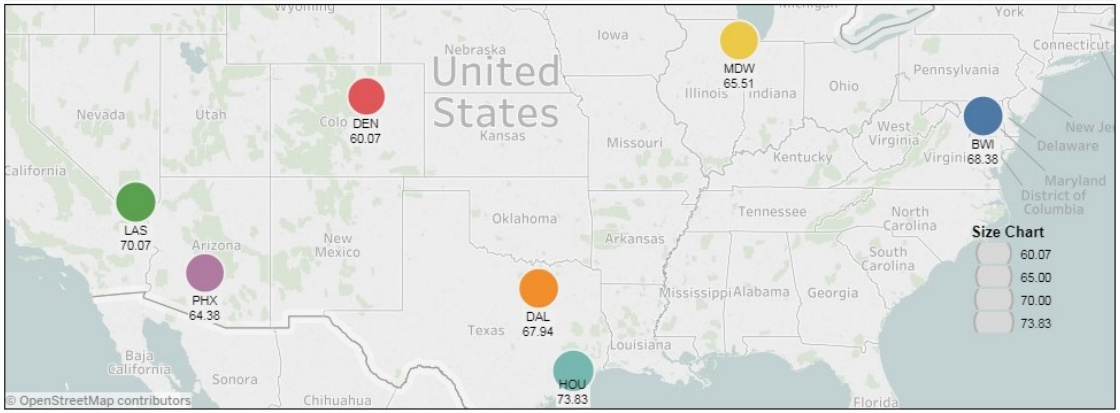
OTP comparison for the last 7 days

Top 10 Factors Explaining Today's Prediction Results

-0.0070 LF_KORF	-0.0087 mon_11	-0.3525 OTP	-0.0165 P12_KDTW	-0.0167 P12_KPHL	-0.0180 SNW_KROC	-0.0091 TMP_KECP	-0.0284 TMP_KMSP	-0.1437 total_flt	-0.0421 WDS_KDEN

Show what factors influence most on the prediction

Station Turn Compliance Prediction



A map view showing station level predictions

DataScience@SMU

Deployment

Diana Shen

Predicting Tomorrow

- When to run?
- When will the data be ready?
- Channel for showing the prediction

Model Maintenance

- Model retraining frequency
- Process of maintaining the model/dashboard

DataScience@SMU

Business Challenges

Diana Shen

Customer Feedback

DataScience@SMU

Future Directions

Diana Shen

Further Expand the Project

- Network level OTP—station level OTP—flight level OTP (flight delays)
- Real decision-driven models
- More models informing customers of potential upcoming repercussion

DataScience@SMU