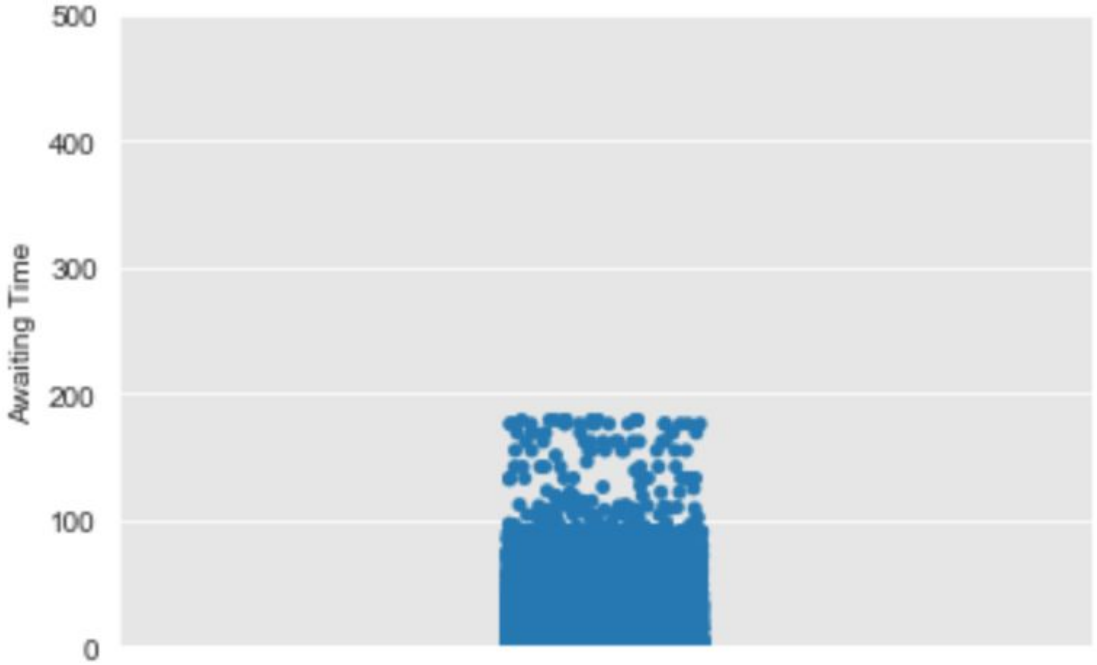
Predicting Medical Appointment No-Shows

Hussein Sajid, Anna Zubova

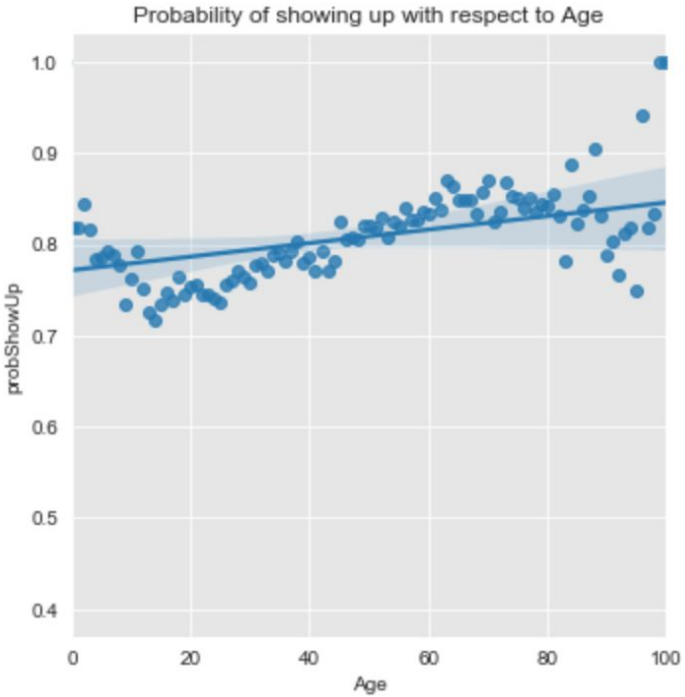
Dataset

- Health care data from Brazil: > 110,000 appointments from year 2016
- Features:
  - Gender
  - Day when appointment was scheduled
  - Appointment Day
  - Age
  - Neighbourhood
  - On welfare or not
  - Medical condition: Hipertension, Diabetes, Alcoholism, Handicap
  - SMS notification: yes or no

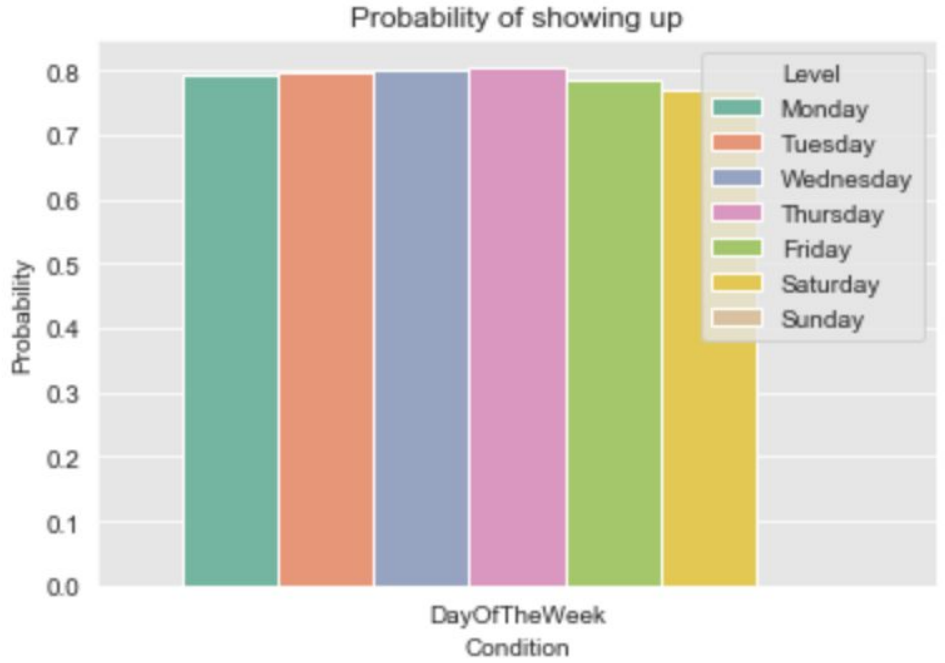
Checking for Outliers in Awaiting Time



Age Feature Probability



Show up Probability



Data cleaning

- Data types
- No null values
- Imbalanced classes of predicted variable: only 20% No-shows
- Transform variables into binary (one-hot-encoding)
- Feature engineering

Baseline model

- No data modifications
- **Algorithm**: Random Forests
- **Features**: Gender, Age, Scholarship, Hipertension, Diabetes, Alcoholism, Handicap, SMS_received
- **Score:** 79%

No-show	Precision	Recall	F1-score
No	0.80	0.99	0.89
Yes	0.32	0.02	0.03

Feature engineering

Create features:

- How many days in advance the appointment was made
- Appointment month
- Appointment day of the week
- Number of prior appointments for each appointment
- Number of prior no-shows for each appointment

Try different models with new features:

model	cv score	f1 0 / 1	precision 0 / 1	recall 0 / 1
decision trees	0.75	0.76 / 0.41	0.89 / 0.30	0.66 / 0.66
random forests	0.71	0.87 / 0.25	0.83 / 0.32	0.90 / 0.20
logistic regression (initial)	0.69	0.84 / 0.33	0.85 / 0.32	0.84 / 0.34
logistic regression (dropped features)	0.69	0.84 / 0.33	0.85 / 0.32	0.83 / 0.35

Logistic regression: most significant features

- Number of previous no-shows (coef = 0.6):
- Appointment in June (coef = - 0.46):
- Received SMS notification (coef = 0.39)

Logistic regression: most significant features

Mario:

- Gender: Male
- Age: 30
- Hipertension: 0
- Alcoholism: 0
- Handicap: 0
- **SMS_received: 0**
- days_in_advance: 14
- **Month_appointment_6: 0**
- Day_of_week_appointment_1: 0
- Day_of_week_appointment_2: 1
- Day_of_week_appointment_3: 0
- Day_of_week_appointment_4: 0
- Number_of_previous_apptms: 0
- **number_of_previous_noshows: 0**

Logistic regression: most significant features

Baseline probability: 46%

Number of previous no-shows: increase from 0 to 1

- Increase of probability from 46% to 61%

Number of previous no-shows: increase from 1 to 2

- Increase of probability from 61% to 74%

Logistic regression: most significant features

Appointment in June :

- Decrease of probability from 74% to 64%

Received SMS notification:

- Increase of probability from 62% to 72%

Thank you!

Questions?