

## Absenteeism at Work

### Business Problem

The online retailer “Amazing Zone” sells various products through its online portal. The products marked “Retailer Fulfilled” are the responsibility of the retailer to provide timely shipments. There are many components in shipment but the majority of last mile delivery is done by delivery boys. This component is important because these delivery boys directly interact with customers while delivering products.

Unplanned absenteeism of these delivery boys from work is a major problem for the online retailer as it leads to delay in delivery fulfillment and customer satisfaction. This further leads to muted sales by the company.

### Observations

As per the data provided, below observations were recorded:

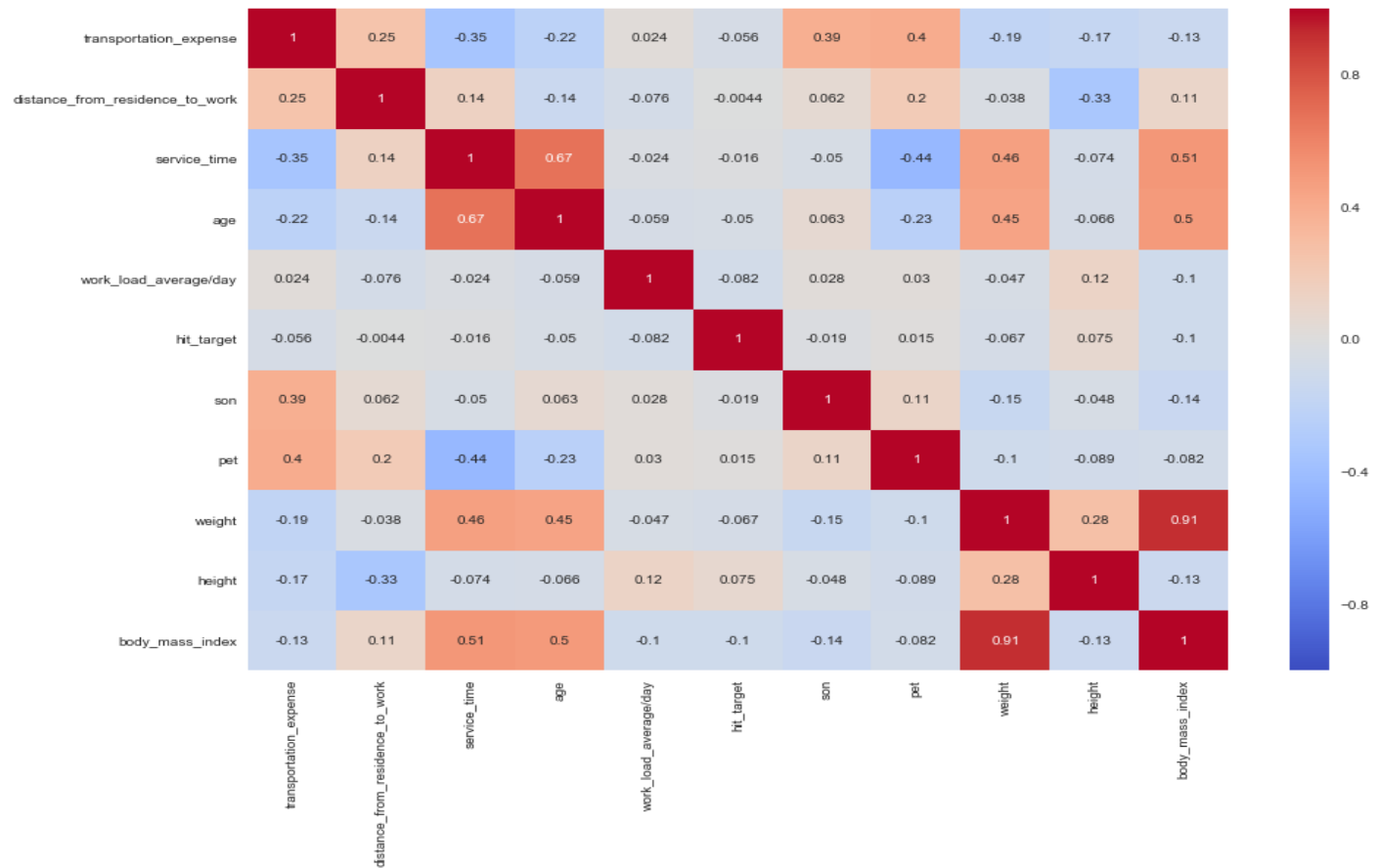
1. There were rows with `reason_for_absence` and `month_of_absence` with 0 value. This had to be treated as null value and were imputed appropriately.
2. There were few `month_of_absence` which overlapped with multiple seasons. This was adjusted such that each month falls into a single season only.
3. `Disciplinary_failure` column was dropped and a new `reason_for_absence` (value = 29) was introduced for this purpose. The `absenteeism_time_in_hours` is 0.
4. Highest absence occurs in March, July and November which coincides with quarter ends and holiday season. March and July are near the quarter ends and hence target deadlines would be very tight. November has a lot of festival days and hence, the absent hours could be more.
5. Most absences occur due to reasons below:
  - a. Diseases of the musculoskeletal system and connective tissue – could be because of bad posture at work, lifting heavy weights of parcel etc.
  - b. Injury, poisoning and certain other consequences of external causes – could be because of accidents while delivering goods
  - c. Medical consultation
  - d. Dental consultation
6. Maximum absenteeism happens on Mondays. This could be so that employees could take long weekends.
7. Maximum absenteeism happens in the lowest level of education but this could be because the data is skewed as last mile delivery boys do not have higher education criteria for employment.
8. More absenteeism are by employees who are social drinkers
9. The dataset has many observations where `absenteeism_time_in_hours` is greater than 24. These observations are removed from dataset as outliers/invalid data.
10. There are few continuous variables which are dependent on each other. Hence, to remove multicollinearity from the dataset, such features are removed.
  - a. `Transportation_expense` is positively correlated with `distance_from_residence_to_work`, hence `transportation_expense` is dropped
  - b. `Service_time` and `age` is positively correlated. Hence, `service_time` is being dropped

c. Body\_mass\_index is calculated from height and weight. Hence, height and weight are being dropped

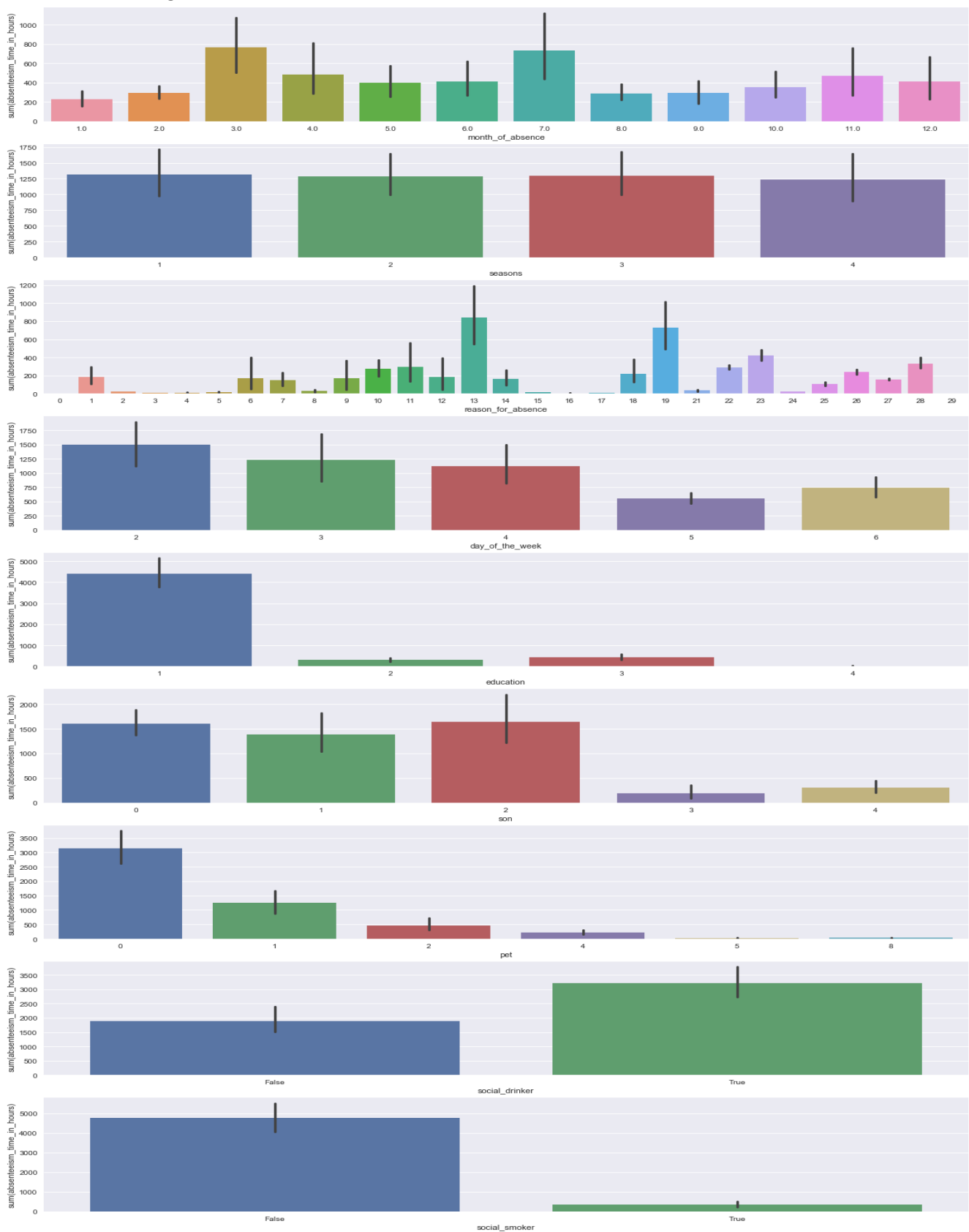
11. Feature Importance is calculated using ExtraTreesClassifier and the threshold is chosen as 0.2 to get 12 features to model the Linear Regression. The threshold is chosen higher to limit the number of features. This leads to a better model explain ability and avoid overfitting model.

## Data Visualizations

Correlation Heat map for all continuous variables to identify feature correlations for multicollinearity



Bar Charts for Categorical Variables vs Absenteeism Time In Hours



## Model Summary and Interpretation

In order to avoid overfitting and to better explain the linear regression model, the below features were chosen:

Features	Coefficient
Intercept	6.59234455518
distance_from_residence_to_work	-0.215232
age	-1.637281
work_load_average/day	0.667181
hit_target	-0.745155
son	0.291016
social_drinker	0.772308
pet	-0.198722
body_mass_index	0.148366
reason_for_absence_22	1.125312
reason_for_absence_23	-3.558712
reason_for_absence_27	-3.780306
reason_for_absence_28	-3.737174
reason_for_absence_29	-6.450236
day_of_the_week_2	-0.011761

Root Mean Squared Error: 2.76088125453

r<sup>2</sup> Score: 0.361457958842

## Suggestions

Below suggestions could be implemented to reduce the number of absenteeism hours from delivery boys:

1. Provide employees with conveyance for travelling to office to reduce their travel time or else provide incentives to reduce their distance from residence to work.
2. Employees with higher age tend to have lesser absenteeism hours. This could be because they have families to take care of and require a permanent job and cannot take a risk of unjustified absence from work leading to their termination.
3. Work load should be adjusted between employees so that absences are controlled.
4. Employees who achieves targets tend to be absent lesser. Such employees should be awarded so that there is a sense of healthy competition between employees to achieve target.
5. People with higher body mass index tend to be more absent. This could be because of associated diseases with obesity.
6. There are many factors from the model which are associated with overall health of employees. Absences due to these factors can be controlled using below suggestions:
  - a. Regular free medical/dental check-ups arranged by the employer
  - b. Free Physiotherapy sessions arranged by the employer
  - c. Proper reasoning should be asked from employees with medical certificate (if needed) for Unjustified Absence

7. Although employees with disciplinary issues (reason\_code=29), do not have absent hours, it should be avoided as such cases bring disrepute to the company. Such cases can be avoided by providing soft skills training to help them to interact with customers.