

## Домашка «Титаник»

1. Зарегистрируйтесь на сайте [www.kaggle.com](http://www.kaggle.com) в конкурсе «Titanic: Machine Learning from Disaster». В работе укажите login, использованный при регистрации.
2. Проанализируйте данные графически и с помощью описательных статистик (среднее, мода, медиана и т.д.)  
Прокомментируйте графики, обратите внимание на количество пропущенных значений.
3. Оцените logit и probit модели.  
Приведите оценки моделей. Какие коэффициенты значимы? Прокомментируйте знак коэффициентов. Посчитайте и сравните предельные эффекты.
4. Оцените random forest и SVM модели.  
Параметры методов подберите с помощью кросс-валидации. Можно применять любые другие подходы, не только random forest и SVM. Другой подход следует описать в тексте.
5. «Если бы я был пассажиром Титаника, то я спасся бы с вероятностью...».  
С помощью логит и пробит моделей постройте 95%-ый доверительный интервал для вероятности своего спасения. Для random forest — только точечный прогноз вероятности, для svm — только прогноз типа «да»/«нет».
6. Подумайте, чем можно заполнить пропущенные значения. Заполните пропущенные значения и заново оцените logit, random forest и svm. Насколько сильно меняется качество оцененных моделей?
7. Сравните все использованные подходы по прогнозной силе на тестовой выборке с сайта. Какой оказался наилучшим?
8. При прогнозировании и расчете предельных эффектов используйте свои фактические пол и возраст, а остальные объясняющие переменные — выбирайте согласно своей фантазии :)
9. Срок сдачи — 30 апреля 2014 года.  
Работа принимается исключительно в печатном виде с применением грамотного программирования R + L<sup>A</sup>T<sub>E</sub>X. Каждый день более поздней сдачи умножает оценку за работу на 0.8. Работа должна представлять слитный текст, код скрывать не нужно. В конце должна быть команда `sessionInfo()`.
10. Популярные ошибки прошлой домашки будут караться со всей строгостью военного времени!  
Цикл заметок про R в помощь <https://github.com/bdemeshev/em301/wiki/R>.