Прогнозирование доходов пользователей социальных сетей.

Меркулов Михаил Владимирович

Студент

Факультет №8 Московский авиационный институт (национальный исследовательский университет), Москва, Россия

E-mail: hungryangry666@gmail.com

Научный руководитель —

Доход пользователей играет значительную роль во многих прикладных задачах. Многие политологические, социологические и маркетинговые модели используют доход как один из ключевых показателей респондента. В данной работе представлен новый способ прогнозирования дохода пользователя, основанный на использовании алгоритма Nod2Vec, для случайного блуждания по графам социальных сетей, и обучении классификатора Random forest по данным блуждания и характеристикам сетевого профиля.

Материалы резюме были получены через публично доступный API ресурса HeadHunter.ru, сетевая информация о сетевых профилях была взята через публично доступный API ресурса VK.com за 2018 год. Для сбора были использованы запросы по заранее составленному словарю профессий: менеджер, программист, бухгалтер, инженер, слесарь и т.д. Публично доступный профиль ресурса HeadHunter.ru не предоставляет информации о контактных данных и имени пользователя, но содержит точные данные об образовании, поле и возрасте, чего достаточно для поиска в социальных сетях. Поскольку поиск по данным характеристикам задает достаточно широкий коридор, в среднем 10-20 страниц на каждый запрос, для омонимичных пользователей было реализовано дополнительное сопоставление фотографии профиля социальной сети и резюме при помощи компонента openface . Предложенный метод имеет высокие характеристики точности, но низкую полноту, удалось сопоставить порядка 3,000 резюме из 10,645. Для полученных профилей были собраны данные, включающие следующие атрибуты: ФИО, дата рождения, пол, возраст, данные об образовании, место жительства, место работы, должность, заработная плата, группы в социальных сетях, друзья и аналогичная информация для друзей.

Полученные профили были полностью анонимизированы, данные об именах были исключены из анализа, данные о дне рождения были преобразованы в возраст. Данные об образовании и городе

проживания были также предобработаны. Университеты прокодированы в соответствии с рейтингом : для университетов из первой десятки ставился в соответствие код - 2, для вузов с 11 по 40 код - 1, для остальных код - 0.

Модель обученная на основе данных случайного блужданият (embedings) и таких признаков как пол, возраст и образование и т.д. при двух-классовой классификации показывает хорошие результаты, до 77.18%. Увеличивая точность базовой модели, обученной без данных случайного блуждания на 5-7%.