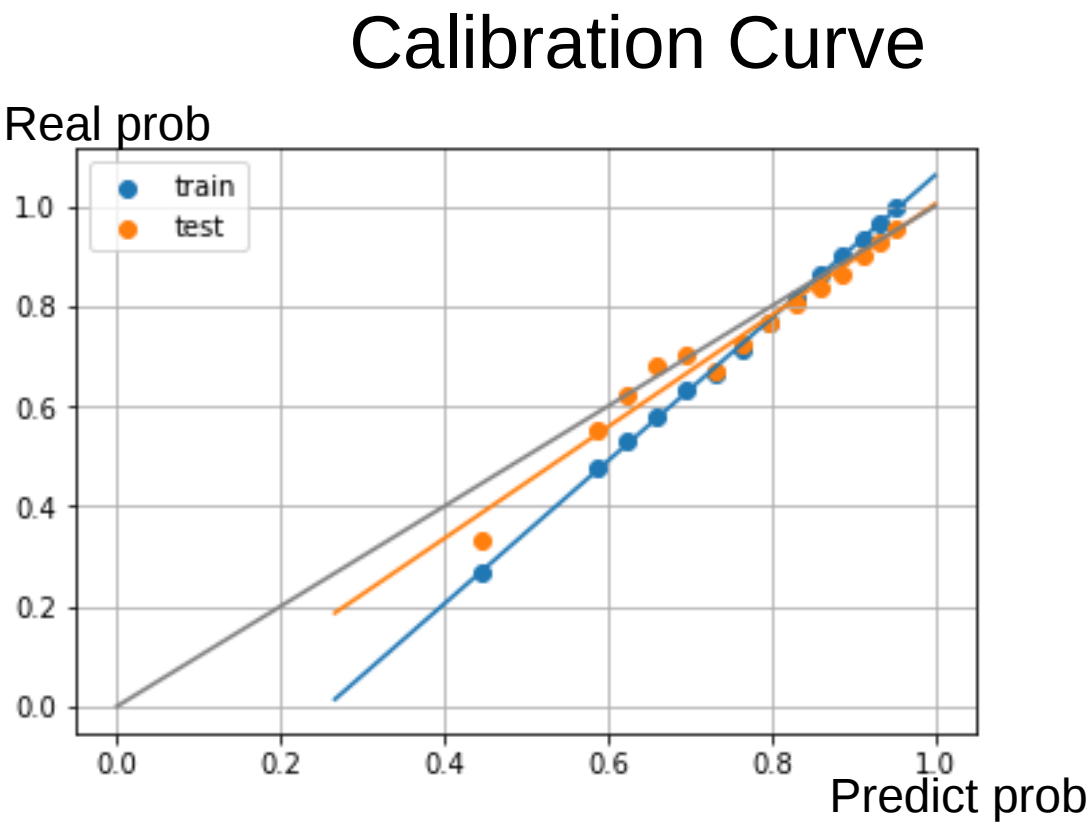
Изучение свойств калибровочных кривых в задаче предсказания CTR. Использование CC для выбора гиперпараметров.



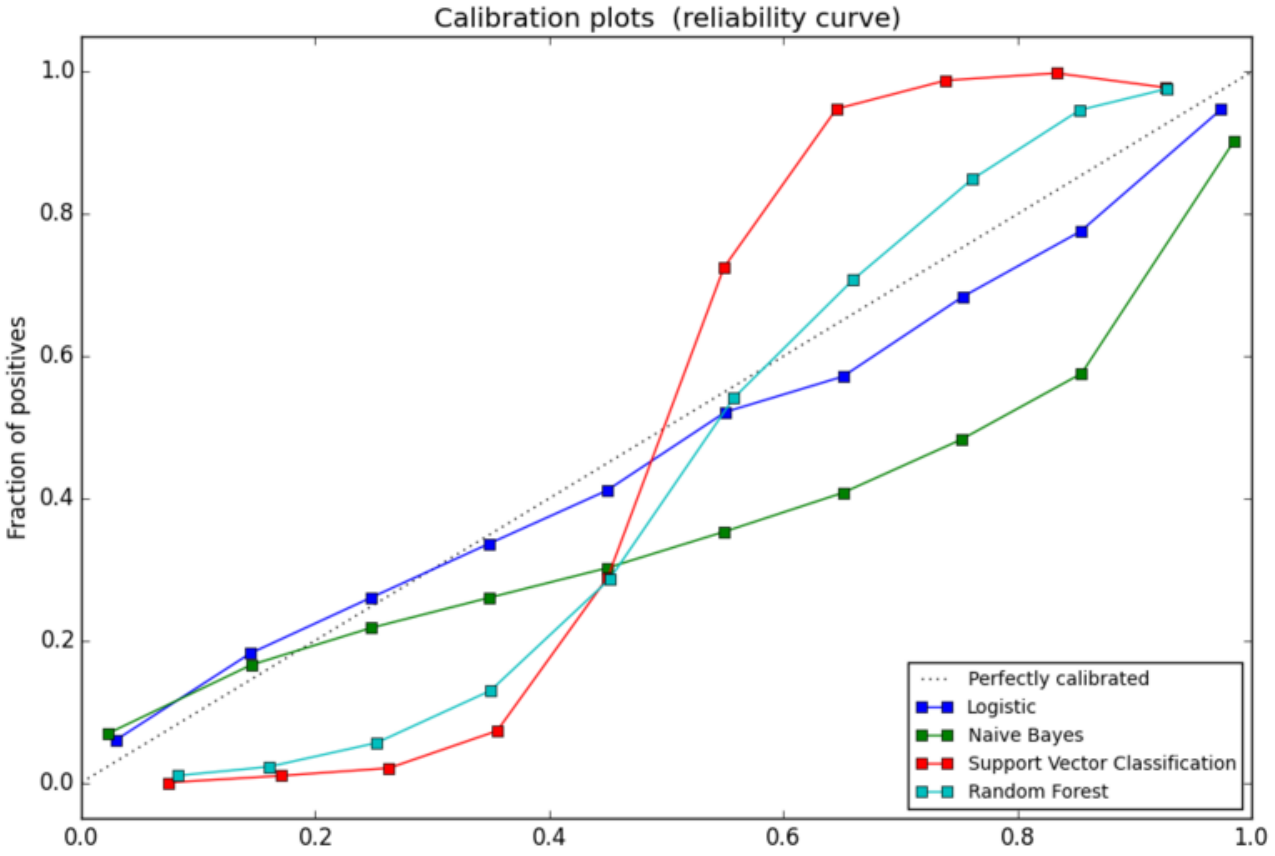
Поповкин Андрей, 4 курс ФИВТ МФТИ
Руководитель: Ворожцов Артем

Введение.
Предсказание CTR.

- **Экономический запрос**
- **Модель классификации**
- **Необходимость предсказания “хорошей” вероятности**
- **Сложное распределение вероятности клика**
- **Огромное количество данных и обширное пространство фичей**
- **Как следствие продолжительное обучение**

Введение.
“Хорошая” вероятность.

- **Калибровочная кривая**

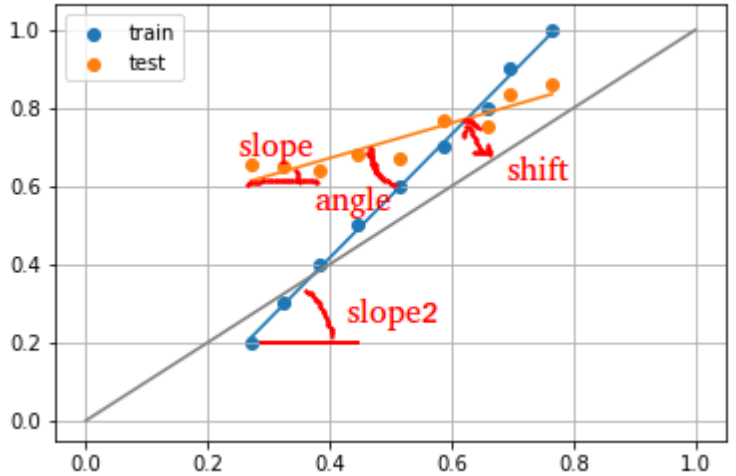


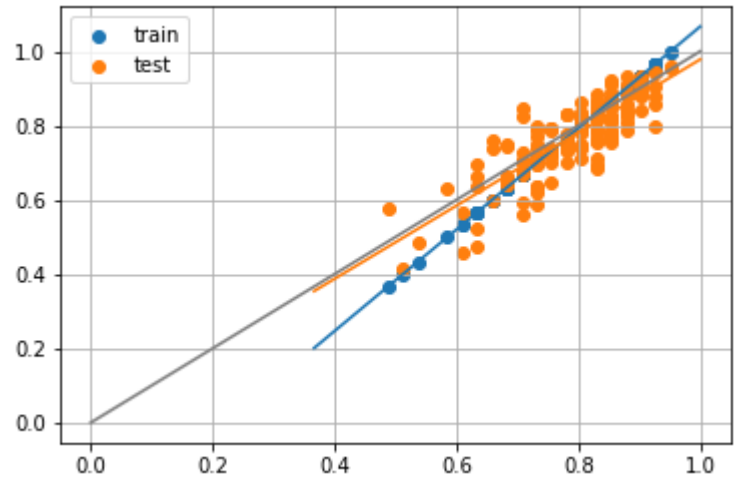
Введение.
Модель.

- **Модель данных**
  - nClasses классов
  - L семплов в каждом из них
  - Вероятность клика для класса семплируется из Beta(a0, b0)
  - Train и test
- **Модель обучения**
  - Баессовский классификатор с prior Beta(a_pr, b_pr)

Введение.
Наблюдения.

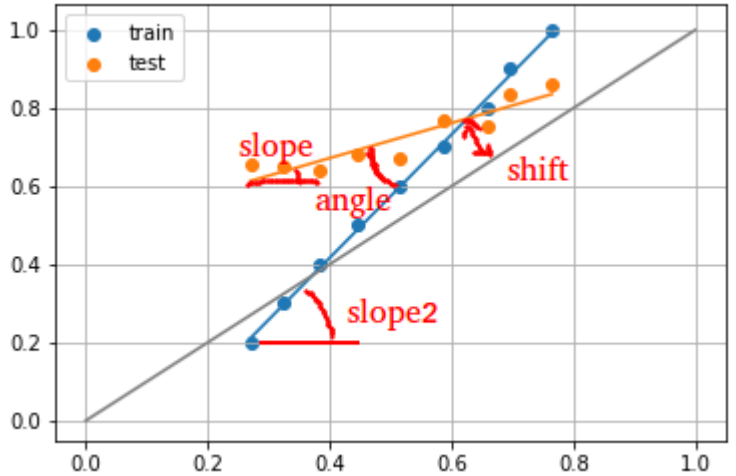
- **Получаются графики ->**
  - Линейные регрессии на точках калибровочных кривых
- **Выделяем характеристики этих прямых**
- **Изучаем их зависимость от параметров модели**
- **Какие характеристики инвариантны относительно каких параметров и почему**

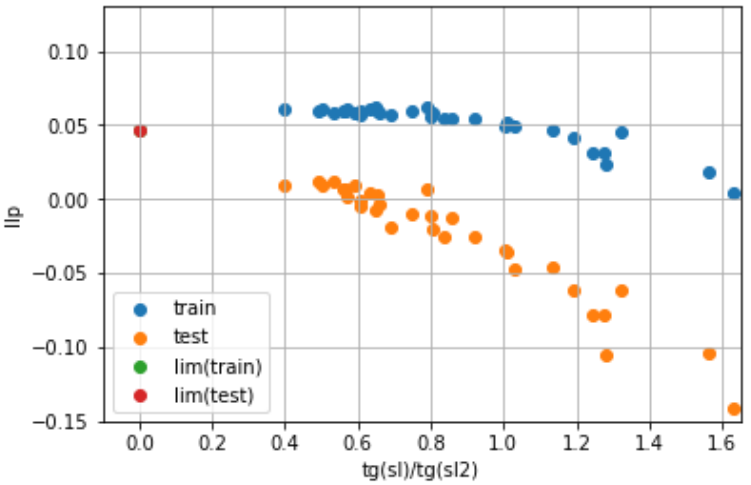




Введение.
Наблюдения.

- **Важный инвариант – tg(slope2) / tg(slope) сохраняется если не менять количество семплов.**
- **Возможно есть непосредственная зависимость LLP от этого отношения**
- **LLP =** $\frac{logLike(model.predict()) - logLike(const)}{\sum(clicks)}$





Цели.

- **Изучение и математическое обоснование указанных свойств.**
- **Определение границ применимости полученных наблюдений: более сложные модели обучения (Catboost, WV) и модели данных.**
- **Потенциальное применение к реальным данным обучения CTR формул.**
- **Идеальным практическим результатом было бы замечание о применимости зависимости LLP от отношения тангенсов к оценке и сравнению моделей на небольшом количестве данных и, как следствие, малом времени обучения.**

Литература

- **Калибровочные кривые являются обьектом исследований, в первую очередь для достижения интерпретируемости выходов моделей, как вероятностей классов.**
- **Существуют исследования, посвященные характерным особенностям калибровочных кривых, которые дают различные модели на задачах классификации.**
- **Про обучение CTR формул написано вообще довольно много, в том числе, есть предложения способов борьбы с малым количеством данных. Что представляет наиболее близкий к данному исследованию кейс.**
- **В целом, распространена идея, что при малом количестве данных, методы максимизирующие правдоподобие слабее чем Байесовский подход.**

План

- **Реализация инструмента исследования описаной модели на языке python.**
- **Фиксация закономерностей, построение графиков, определение наиболее интересных для дальнейшего изучения.**
- **Повторение зависимостей на других моделях обучения.**
- **Математическое обоснование.**
- **Усложнение модели данных.**
- **Адаптация существующих датасетов обучения CTR формул.**
- **Попытаться применить полученные закономерности к мета-оптимизации.**

Прогресс

- **Переписаны на python (исходный код на Matematica) инструменты для исследования тестовой задачи с Баессовским классификатором.**
- **Воспроизведены основные результаты.**
- **Начато исследование для модели Catboost.**

Начато практическое изучение фреймворка hyperopt и его версии для работы с VW… А также подходов о оптимизации гиперпараметров.