Classificação das Tipologias de Imóveis

Uma Aplicação do *Gradient Boosting* para o Município de Belo Horizonte-MG

Alan Leal

FEA/USP

11/04/2022

Um imóvel está sendo negociado no mercado próprio ou no mercado de terrenos?

- Os imóveis são negociados em mercado próprio ou mercado de terrenos? Isto é, seu preço está definido principalmente por sua construção ou pelo solo sob o qual ele se encontra construído?

Um imóvel está sendo negociado no mercado próprio ou no mercado de terrenos?

- Os imóveis são negociados em mercado próprio ou mercado de terrenos? Isto é, seu preço está definido principalmente por sua construção ou pelo solo sob o qual ele se encontra construído?
- O que em termos gerais poderia caracterizar o mercado de terrenos e o mercado imobiliário próprio de cada imóvel?

Um imóvel está sendo negociado no mercado próprio ou no mercado de terrenos?

- Os imóveis são negociados em mercado próprio ou mercado de terrenos? Isto é, seu preço está definido principalmente por sua construção ou pelo solo sob o qual ele se encontra construído?
- O que em termos gerais poderia caracterizar o mercado de terrenos e o mercado imobiliário próprio de cada imóvel?
- Como identificar na prática e em massa os imóveis que são negociados em mercado próprio ou no mercado de terrenos?

Um imóvel está sendo negociado no mercado próprio ou no mercado de terrenos?

- Os imóveis são negociados em mercado próprio ou mercado de terrenos? Isto é, seu preço está definido principalmente por sua construção ou pelo solo sob o qual ele se encontra construído?
- O que em termos gerais poderia caracterizar o mercado de terrenos e o mercado imobiliário próprio de cada imóvel?
- Como identificar na prática e em massa os imóveis que são negociados em mercado próprio ou no mercado de terrenos?
- É viável realizar essa classificação imóvel a imóvel?
- Principais métodos e possibilidades

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT
- Modelos de aprendizagem de máquina supervisionado

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT
- Modelos de aprendizagem de máquina supervisionado
  1. *SVM*

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT
- Modelos de aprendizagem de máquina supervisionado
  1. *SVM*
  2. Árvore de decisão

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT
- Modelos de aprendizagem de máquina supervisionado
  1. *SVM*
  2. Árvore de decisão
  3. *Random forests*

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT
- Modelos de aprendizagem de máquina supervisionado
  1. *SVM*
  2. Árvore de decisão
  3. *Random forests*
  4. Vizinhos mais próximos

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT
- Modelos de aprendizagem de máquina supervisionado
  1. *SVM*
  2. Árvore de decisão
  3. *Random forests*
  4. Vizinhos mais próximos
  5. *Bagged trees*

Métodos de classificação

- Modelos econométricos de resposta binária: LOGIT e PROBIT
- Modelos de aprendizagem de máquina supervisionado
  1. *SVM*
  2. Árvore de decisão
  3. *Random forests*
  4. Vizinhos mais próximos
  5. *Bagged trees*

Trade-offs dos métodos de classificação

- *Trade-off* entre interpretabilidade e previsibilidade

Trade-offs dos métodos de classificação

- *Trade-off* entre interpretabilidade e previsibilidade
- Métodos de classificação caixa-preta

Trade-offs dos métodos de classificação

- *Trade-off* entre interpretabilidade e previsibilidade
- Métodos de classificação caixa-preta
- Implementação do método *bagged trees* via algoritmo do *Gradient Boosting*

Trade-offs dos métodos de classificação

- *Trade-off* entre interpretabilidade e previsibilidade
- Métodos de classificação caixa-preta
- Implementação do método *bagged trees* via algoritmo do *Gradient Boosting*

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
     1. Classifique os dados com base nos pesos atribuídos

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
     1. Classifique os dados com base nos pesos atribuídos
     2. Meça o erro de classificação

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
     1. Classifique os dados com base nos pesos atribuídos
     2. Meça o erro de classificação
     3. Calcule o log-erro da sua tarefa de classificação atual

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
     1. Classifique os dados com base nos pesos atribuídos
     2. Meça o erro de classificação
     3. Calcule o log-erro da sua tarefa de classificação atual
     4. Use o log-erro (ou outra métrica de erro) para atualizar seus pesos, de modo a dar maior peso na tarefa de classificação às amostras erroneamente classificadas

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
     1. Classifique os dados com base nos pesos atribuídos
     2. Meça o erro de classificação
     3. Calcule o log-erro da sua tarefa de classificação atual
     4. Use o log-erro (ou outra métrica de erro) para atualizar seus pesos, de modo a dar maior peso na tarefa de classificação às amostras erroneamente classificadas
     5. Pule para a n+1 classificação

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
     1. Classifique os dados com base nos pesos atribuídos
     2. Meça o erro de classificação
     3. Calcule o log-erro da sua tarefa de classificação atual
     4. Use o log-erro (ou outra métrica de erro) para atualizar seus pesos, de modo a dar maior peso na tarefa de classificação às amostras erroneamente classificadas
     5. Pule para a n+1 classificação
  2. Retorne a árvore de classificação

O algoritmo Gradient Boosting para a classificação dos imóveis

- Intuição por trás do *Gradient Boosting*
- Implementação do algoritmo *ADABoost*:
  1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
     1. Classifique os dados com base nos pesos atribuídos
     2. Meça o erro de classificação
     3. Calcule o log-erro da sua tarefa de classificação atual
     4. Use o log-erro (ou outra métrica de erro) para atualizar seus pesos, de modo a dar maior peso na tarefa de classificação às amostras erroneamente classificadas
     5. Pule para a n+1 classificação
  2. Retorne a árvore de classificação

Implementando na prática o Gradient Boosting

- Variável dependente: intuição e construção

Implementando na prática o Gradient Boosting

- Variável dependente: intuição e construção
  1. Junção do Índice Cadastral da base do ITBI à base de projetos

Implementando na prática o Gradient Boosting

- Variável dependente: intuição e construção
  1. Junção do Índice Cadastral da base do ITBI à base de projetos
  2. Espera-se que imóveis associados a algum projeto tenham maior probabilidade de estarem sendo negociados no mercado imobiliário de terrenos e não exatamente no mercado próprio de sua tipologia

Implementando na prática o Gradient Boosting

- Variável dependente: intuição e construção
  1. Junção do Índice Cadastral da base do ITBI à base de projetos
  2. Espera-se que imóveis associados a algum projeto tenham maior probabilidade de estarem sendo negociados no mercado imobiliário de terrenos e não exatamente no mercado próprio de sua tipologia
  3. Dessa forma, a variável de classificação no modelo (variável *match*) consiste numa variável binária assumindo o valor 1, caso o imóvel do ITBI esteja associado a algum projeto, e 0, caso contrário

Implementando na prática o Gradient Boosting

- Variável dependente: intuição e construção
  1. Junção do Índice Cadastral da base do ITBI à base de projetos
  2. Espera-se que imóveis associados a algum projeto tenham maior probabilidade de estarem sendo negociados no mercado imobiliário de terrenos e não exatamente no mercado próprio de sua tipologia
  3. Dessa forma, a variável de classificação no modelo (variável *match*) consiste numa variável binária assumindo o valor 1, caso o imóvel do ITBI esteja associado a algum projeto, e 0, caso contrário
  4. Previsão e uso futuro do modelo para novos dados imobiliários do ITBI

Implementando na prática o Gradient Boosting

- Variável dependente: intuição e construção
  1. Junção do Índice Cadastral da base do ITBI à base de projetos
  2. Espera-se que imóveis associados a algum projeto tenham maior probabilidade de estarem sendo negociados no mercado imobiliário de terrenos e não exatamente no mercado próprio de sua tipologia
  3. Dessa forma, a variável de classificação no modelo (variável *match*) consiste numa variável binária assumindo o valor 1, caso o imóvel do ITBI esteja associado a algum projeto, e 0, caso contrário
  4. Previsão e uso futuro do modelo para novos dados imobiliários do ITBI

Algumas variáveis utilizadas exercício empírico

- Variáveis:

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais
  10. Distâncias do imóvel às escolas mais próximas, públicas e privadas

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais
  10. Distâncias do imóvel às escolas mais próximas, públicas e privadas
  11. Número de bancos, farmácias e indústrias no bairro do imóvel

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais
  10. Distâncias do imóvel às escolas mais próximas, públicas e privadas
  11. Número de bancos, farmácias e indústrias no bairro do imóvel
  12. Média dos imóveis vizinhos que são apartamentos (wAP)

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais
  10. Distâncias do imóvel às escolas mais próximas, públicas e privadas
  11. Número de bancos, farmácias e indústrias no bairro do imóvel
  12. Média dos imóveis vizinhos que são apartamentos (wAP)
  13. Média da idade dos imóveis vizinhos (widade)

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais
  10. Distâncias do imóvel às escolas mais próximas, públicas e privadas
  11. Número de bancos, farmácias e indústrias no bairro do imóvel
  12. Média dos imóveis vizinhos que são apartamentos (wAP)
  13. Média da idade dos imóveis vizinhos (widade)
  14. Interação entre wAP e widade

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais
  10. Distâncias do imóvel às escolas mais próximas, públicas e privadas
  11. Número de bancos, farmácias e indústrias no bairro do imóvel
  12. Média dos imóveis vizinhos que são apartamentos (wAP)
  13. Média da idade dos imóveis vizinhos (widade)
  14. Interação entre wAP e widade
  15. Dummy_app: *dummy* se o imóvel está ou não em área de proteção ambiental

Algumas variáveis utilizadas exercício empírico

- Variáveis:
  1. Valor declarado do imóvel
  2. Ano de transação
  3. Padrão de acabamento (sob a forma de cinco *dummies*)
  4. Área construída adquirida
  5. Área do terreno adquirida
  6. Idade do imóvel
  7. IQVU
  8. Distância do imóvel ao hospital mais próximo
  9. Distância do imóvel a praças regionais
  10. Distâncias do imóvel às escolas mais próximas, públicas e privadas
  11. Número de bancos, farmácias e indústrias no bairro do imóvel
  12. Média dos imóveis vizinhos que são apartamentos (wAP)
  13. Média da idade dos imóveis vizinhos (widade)
  14. Interação entre wAP e widade
  15. Dummy_app: *dummy* se o imóvel está ou não em área de proteção ambiental

Implementação prática

- Divisão da amostra

Implementação prática

- Divisão da amostra
- Overfitting/Underfitting: como impedir que um deles ocorra

Implementação prática

- Divisão da amostra
- Overfitting/Underfitting: como impedir que um deles ocorra
- Implementação em várias linguagens: R, Python, MATLAB, dentre outras

Implementação prática

- Divisão da amostra
- Overfitting/Underfitting: como impedir que um deles ocorra
- Implementação em várias linguagens: R, Python, MATLAB, dentre outras

Métricas de análise do ajuste do modelo

As seguintes métricas são utilizadas como forma de validação do método utilizado

Métrica	Acurácia	AUC	*Matthews correlation coefficient*
Fórmula	$\frac{C}{N}$	Área sobre a curva ROC	$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Significado	C é o número de casos corretos, enquanto N é o número total de casos classificados. A acurácia mede assim a média de acertos do modelo, sem distinção entre casos positivos e negativos	A curva ROC (*receiver operating characteristic curve*) plota a sensitividade contra a taxa de falsos positivos ($\frac{FP}{FP+TN}$). Quanto maior a área sob a curva ROC, maior a precisão do modelo. Em termos práticos, quanto maior o valor de AUC (*area under curve*), maior a chance de que um valor positivo aleatório tenha maior probabilidade predita do que um valor negativo aleatório	TP, TN, FP e FN são, respectivamente, os verdadeiros positivos, verdadeiros negativos, os falsos positivos e falsos negativos, obtidos via matriz de confusão. *Matthews correlation coefficient* mede o acerto do modelo de classificação levando em conta o possível desbalanceamento de classes presentes no modelo. Desse modo, ele é mais confiável que a acurácia

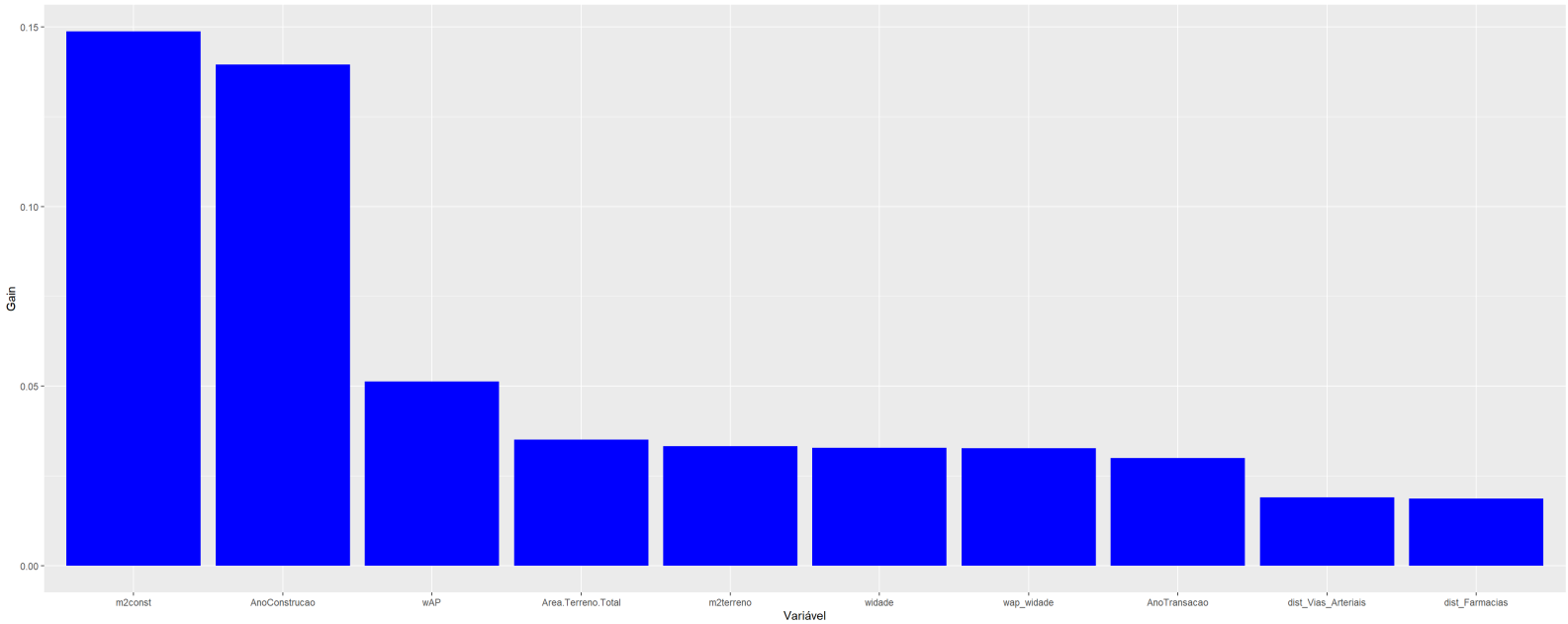
Resultados

Métricas de ajuste do modelo de classificação out-of-sample

Métrica/Tipologia	Casas	Barracões	Galpões
Acurácia	0,9516	0,9381	0,9619
AUC (Área sob a curva ROC)	0,9263	0,8964	0,9475
Coeficiente de correlação de Matthews	0,8686	0,8164	0,9058
Número de observações na amostra de treinamento	14847	1357	1103

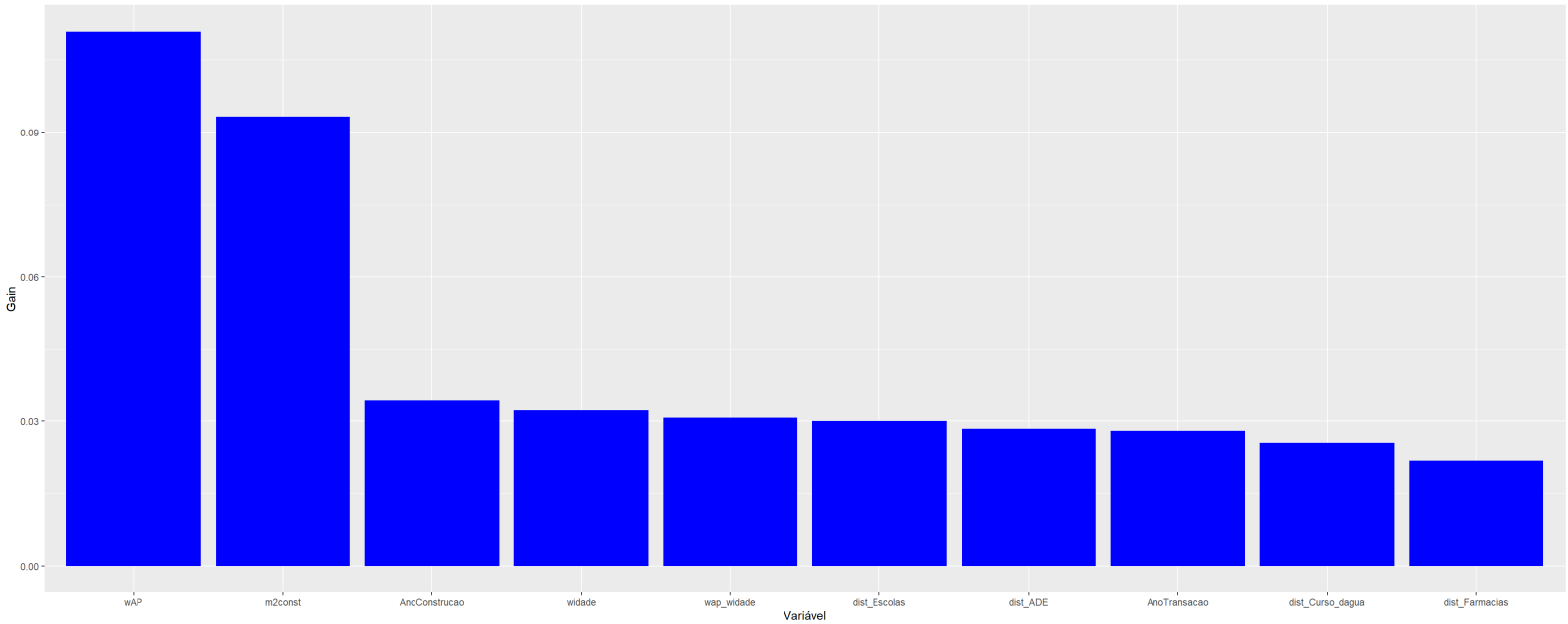
Resultados

Figure: Matriz de ganhos - Casas



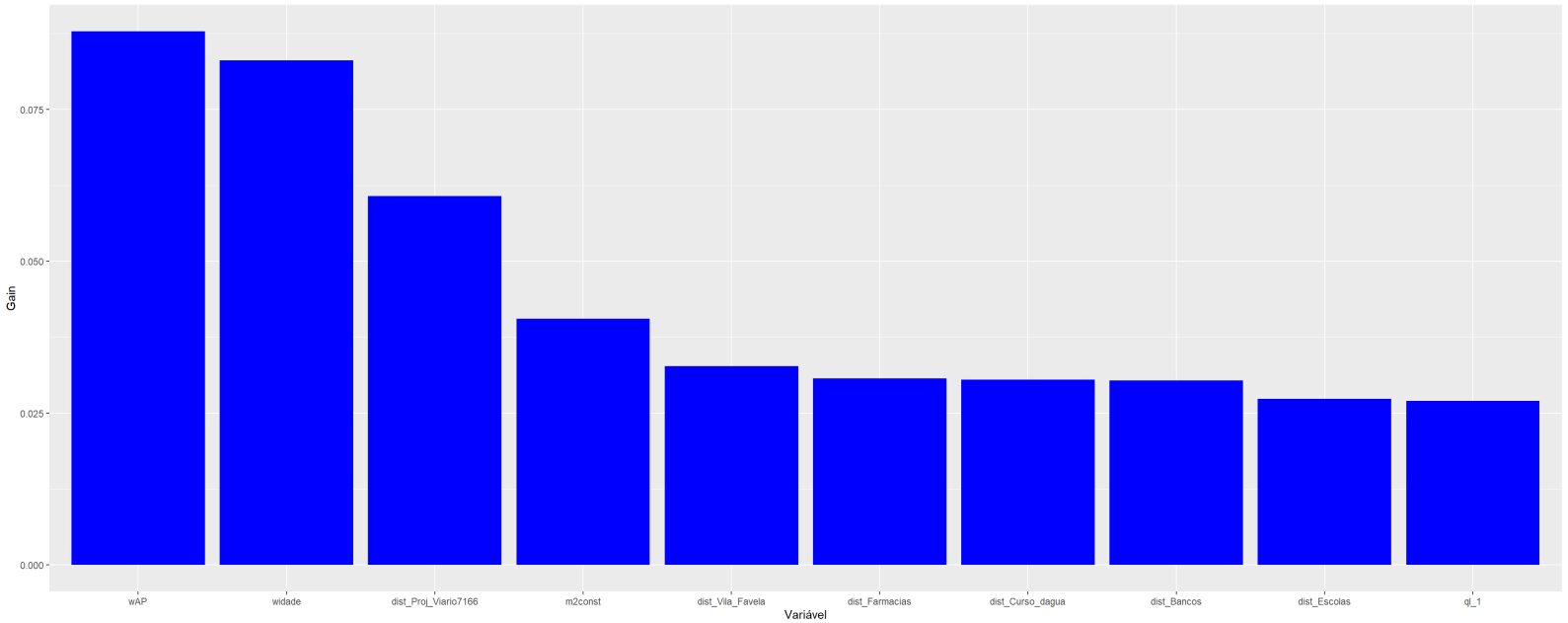
Resultados

Figure: Matriz de ganhos - Barracões



Resultados

Figure: Matriz de ganhos - Galpões



Erro em 5% da classificação: identificação e correção

- A medida de acurácia exibida para o modelo de classificação das três tipologias indicou que cerca de 5% dos imóveis foram classificados erroneamente. Como identificar e corrigir em massa esses erros, ou seja, como implementar mecanismos de correção que sejam adequados e demandem a menor quantidade possível de trabalho humano de checar imóvel a imóvel?
- Dado que o output do modelo de classificação é indicar em qual modelo de preços (próprio ou de terrenos) o imóvel está sendo majoritariamente precificado, duas alternativas são bem naturais, quais sejam:
  1. Comparar o preço estimado no mercado para o qual ele foi classificado e no outro mercado o qual ele não foi classificado
  2. Comparar o preço do imóvel no mercado para o qual ele foi classificado e estimativa prévia do órgão público

Erro em 5% da classificação: mecanismos assimétricos de correção

- Erros para baixo e para cima nos preços advindos do modelo de classficação:
  1. Subestimação do preço do imóvel: uma comparação do preço estimado usando o resultado do modelo de classificação e os preços previamente estimados pelo órgão público fornecem uma primeira regra de decisão ao uso desse novo preço estimado.
  2. Superestimação do preço do imóvel: a mesma comparação anterior aqui é válida, contudo nessa situação há incentivos para o contribuinte questionar o valor estimado do mercado de seu imóvel caso ele esteja muito discrepante do estimado pelo modelo de preços hedônicos.

Erro em 5% da classificação: mecanismos assimétricos de correção

- Na prática, portanto, a subestimação do preço do imóvel a partir de um viés introduzido pelo modelo de classificação tem menos mecanismos externos de correção do que uma superestimação do preço do imóvel. Identificar apropriadamente esses dois casos apresenta desafios, mas ainda assim justificam o ganho de processamento em termos de que 95% se encontram bem classificados e apropriadamente precificados pela prefeitura.

Principais ganhos e limitações do método

- O modelo de classificação utilizando o método do *gradient boosting* permite classificações acertadas *out-of-sample* da ordem de 95%. Para milhares de observações, isso oferece economia do funcionalismo público, além de fazer uso de informações diversas disponíveis de uma forma matematicamente consistente.
- Uma limitação do modelo consiste ainda na discricionariedade utilizada na identificação de imóveis mal-classificados cujas estimativas de preços não vão condizer com o observado na prática. Superestimação de preços fornece incentivos de questionamentos pelo contribuinte, enquanto subestimação de preços dos imóveis não necessariamente fornece esse tipo de questionamento pelo contribuinte. Uso de preços previamente estimados ou de um modelo de preços na outra classificação de imóveis fornecem preços alternativos para aquele estimado com o resultado do modelo de classificação e podem explicitar mais facilmente os erros de classificação.

Referências Bibliográficas

- Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.
- Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.